Fundamentals of Mathematics

THE MACMILLAN COMPANY
NEW YORK · BOSTON · CHICAGO · DALLAS
ATLANTA · SAN FRANCISCO

MACMILLAN AND CO., LIMITED
LONDON · BOMBAY · CALCUTTA · MADRAS
MELBOURNE

THE MACMILLAN COMPANY
OF CANADA, LIMITED
TORONTO

FUNDAMENTALS OF MATHEMATICS

BY

M. RICHARDSON, Ph.D.

Instructor in Mathematics
Brooklyn College

NEW YORK · 1941

The Macmillan Company

— PRINTED IN THE UNITED STATES OF AMERICA —

Published May, 1941

PREFACE

The place of mathematics in a liberal education has been the subject of much discussion in recent years. It has never been doubted that students of the sciences need intensive training in mathematical techniques. But it is more and more generally admitted that students of the arts and social sciences have little need for such technical skill, save for exceptional cases such as the few who want to use and understand statistical methods. What should be given to students of the arts and social sciences, in what is usually their last year of mathematics? The customary freshman course with emphasis on further memorized and regurgitated techniques seems to be far from the best way to use this educational opportunity. A solution of this problem must rest, in my opinion, on two hypotheses of whose truth I am firmly convinced, namely that mathematics has something of value to offer these students, and that the students are not necessarily unintelligent whether or not they are proficient in the routine manipulations of their high school algebra. This book is intended primarily to provide a sound, suitable, and elastic course for the class of students mentioned; however, considerations of the kind discussed here would also be beneficial for students of the sciences who are commonly and naively expected to acquire an understanding of fundamental concepts by osmosis.

The principal objectives of this book are to give the student:

(1) An appreciation of the natural origin and evolutionary growth of the basic mathematical ideas from antiquity to the present;

(2) A critical logical attitude, and a wholesome respect for correct reasoning, precise definitions, and clear grasp of underlying assumptions;

(3) An understanding of the role of mathematics as one of the major branches of human endeavor, and its relations with other branches of the accumulated wisdom of the human race;

(4) A discussion of some of the simpler important problems of pure mathematics and its applications, including some which often come to the attention of the educated layman and cause him needless confusion;

(5) An understanding of the nature and practical importance of postulational thinking.

While it is assumed that the student has had some previous acquaintance with elementary algebra and plane geometry, almost no accurate recollection of the details of these subjects is prerequisite for this book. There is little use in grumbling about the average student's lack of skill in mechanical routines or in rushing him into further ill-understood techniques. Instead, we accept the student and his preparation as incontrovertible data and begin by discussing the reasonable character of the elementary mathematics which he formerly knew largely by rote. After a brief introduction to the essential logical ideas which are fundamental to any appreciation of mathematics, the evolution of the number system and the essentials of elementary algebra are discussed from a mature and reasonable point of view. After this foundation is laid, the book provides an elementary but critical introduction to several of the most important branches of modern mathematics, without, however, pushing any chapter so far as to strain the student's technical equipment. Applications are discussed throughout, but the discussion is restricted to applications which are within the student's grasp rather than to allow it to degenerate into a Sunday Supplement article on the Wonders of Science.

The names of some of the topics taken up may at first glance produce the feeling that the book is too "hard" for freshmen; this feeling, which would have been shared by the author before he began experimenting with such material in his classes, will be partially dissipated by a careful examination of the manner in which these topics are treated, and will, I believe, be obliterated completely by an actual trial in the classroom. I wish to assert as strongly as possible that the early fundamental paragraphs of a so-called "advanced" subject, presented at the proper level, are often easier to grasp as well as more important and more interesting than the later technically complicated paragraphs of what has traditionally passed for an "elementary" subject.

At Brooklyn College, some fifteen instructors used this book in preliminary editions with about 1800 students—entirely unselected—through four semesters. The experience of this large group has confirmed my belief that the subject matter used here is more intelligible, more useful, more appealing, and more appropriate for the freshman student of arts and social sciences than more traditional freshman topics like determinants, synthetic division, complicated trigonometric identities, etc. The fundamental ideas of mathematics constitute a major contribution to human thought and, as such, belong in any liberal education. They can be grasped by freshmen if they are permitted to emerge gradually from simple concrete situations with the aid of common sense as they often did in the actual history of the subject. The notion that these ideas can be grasped only at the graduate school level is pure myth. Moreover, they can be given to freshmen without introducing many complicated skills and techniques.

Of course, most teachers try to discuss fundamentals even in traditional courses, but time is short and is often spent in teaching the students "how to do" problems which they will never meet again in their lives. With students of the arts and social sciences, time may be obtained by lightening the burden of technical achievement. However, a course in the appreciation or philosophy of mathematics for students who have no acquaintance with mathematics would be just as worthless as a course in the appreciation of music for students who have never heard any. Talk is certainly not enough; the student should come to grips with genuine (not necessarily traditional) mathematics. This book is not merely an expository essay, nor an attempt to provide a sugar-coated and worthless course for incompetents. It is an attempt to give the student genuine training in the direction of the five objectives listed above rather than merely in routine techniques.

The author has intended to present a course in mathematics which will emphasize the distinction between familiarity and understanding, between logical proof and routine manipulation, between a critical attitude of mind and habitual unquestioning belief, between scientific knowledge and both encyclopedic collections of facts and mere opinion and conjecture, and which

will give the student a wholesome appreciation of the nature and importance of mathematics.

"Teachability" has been borne in mind constantly in connection with the choice and treatment of material. While the main emphasis is on reasoning and ideas, the book is constructed on broad historical lines. The author has attempted to include enough historical and biographical remarks to give the student a feeling for the evolutionary growth of the subject in response to human needs, for the fact that its progress is due to the efforts of human beings, and for the fact that it is still a living subject at which living human beings work. However, strict chronology is often sacrificed in the interests of logical presentation. Neither time nor our purpose would permit an attempt to follow all the blundering efforts of the human race to develop a satisfactory mathematics. As for the puzzle-motive, strong and universal urge though it is, I believe it can be easily overemphasized; in fact, students are entirely too prone to regard mathematics as a branch of parlor magic, or, at best, as a collection of tricks for the solution of catchy problems which are possibly useful to someone else. The attempt is made to present a critical treatment of mathematical ideas *at the student's level;* complete logical rigor, according to present standards, is to be neither expected nor desired at this level. In fact, one of the most puzzling tasks that has confronted the author has been that of deciding what logical inadequacies should or should not be permitted. Although this is largely a matter of taste, the author does not feel obliged to apologize for not presenting negative numbers and rational numbers as classes of equivalent ordered pairs of natural numbers, for example, or for not proving that the sum of fractions, for instance, is independent of the choice of denominator. It will of course be easy for the instructor to modify the amount of "rigor" according to his taste or the needs of any particular class; if less rigor is desired, this can be achieved simply by omission.

If a "survey" course is to be more than a miscellany, some unifying theme is required. Previous attempts to use the function concept, graphical representation, etc., as the unifying theme have met with varying degrees of success. I believe that

no means complete, and contains both books for further serious study and books for supplementary reading.

I wish to express here my gratitude to the many colleagues who kindly consented to experiment with preliminary versions of this text and who offered helpful suggestions and encouragement. I am particularly indebted to Professor H. F. MacNeish for the elegant example of a finite "euclidean" geometry in Chapter XVII; to Professor W. Prenowitz for communicating to me a sketch of the intuitively simple proofs of the theorems of Lobachevskian geometry in Chapter XVI; to Professors S. Borofsky, L. S. Kennison, W. Prenowitz, Drs. A. W. Landers, J. Singer, and J. Wolfe, and to Professors B. P. Gill and E. L. Post of City College, for reading parts of the manuscript critically; to Professor C. B. Boyer for checking some of the historical comments; to Professor A. Church of Princeton University for some helpful remarks concerning recent work on consistency proofs; and to Professor C. A. Hutchinson of the University of Colorado for his careful reading of the proof-sheets. I am also indebted to Dr. Singer for his assistance with illustrations; to Professor E. T. Bell of the California Institute of Technology for permission to use portraits from his *Men of Mathematics;* to Professor D. C. Miller of the Case School of Applied Science for permission to use the illustration in Fig. 204 taken from his *The Science of Musical Sounds;* to the editors of *Scripta Mathematica* for permission to use portraits from Professor D. E. Smith's portfolio of famous mathematicians; to the Burroughs Adding Machine Company for permission to use the illustrations in Figs. 40 and 41; and to the Keuffel and Esser Company for permission to use the illustrations in Figs. 42, 43, and 172. I wish also to thank the Macmillan Company for their cooperation and efficiency.

M. R.

BROOKLYN, NEW YORK
April, 1941

TABLE OF CONTENTS

CHAPTER VI. THE ALGEBRA OF LOGIC, AND OTHER ALGEBRAS

CHAPTER VII. ARITHMETIC, EXPONENTS, AND LOGARITHMS

CHAPTER VIII. IMPOSSIBILITIES AND UNSOLVED PROBLEMS

CHAPTER IX. ANALYTIC GEOMETRY

CHAPTER X. FUNCTIONS

CHAPTER XI. LIMITS AND THE CALCULUS

Fundamentals of Mathematics

Chapter I

INTRODUCTION

1. Aims and program of the book. Whatever a liberal education may be, it should surely involve some acquaintance with the principal achievements of human endeavor. Certainly, the search for truth is an important branch of human endeavor, and in it mathematics has played a major rôle. What is mathematics? What is its rôle and what are its relations with other branches of knowledge? Why is it important? How did it arise and develop? These are some of the questions we shall discuss here.

Since, in your high school course, you studied mathematics for at least two years, it would seem fair to ask *you* "What is mathematics?" Or at least, "What are its outstanding characteristics?" Most students answer by saying something about the study of useful relations among numbers, quantities, measurements of geometric figures, and methods of solution of problems concerning these things. Even some dictionaries still give such definitions. Indeed in the student's previous experience with mathematics, it is true that the *objects* which he studied were numbers and geometric figures. But let us think rather of the distinctive characteristics of the *methods* by which these things were studied as compared with other subjects. You will remember that in geometry you proved assertions logically, one after another, on the basis of certain fundamental assertions (called *axioms*, *assumptions*, or *postulates*) which were taken for granted. Most of your education, aside from this, was a matter of memory and acceptance of authority as it must be, for example, in the elementary study of languages. In fact, while the idea of logical proof was brought home to you strongly in geometry, you may well be unable to recall anything of the sort in algebra. This is so because, in algebra, you were probably taught to solve certain types of problems according to prescribed rules which in-

volved moving letters and numbers around the paper or blackboard in peculiar ways. Unfortunately "transposition," "cancellation," etc., meant very little, to some of you, except a curious collection of rules of thumb, like a glorified game of tick tack toe, whose correctness was justified not by reasoning but by the teacher's authority. Can you, for example, justify logically the queer maneuvers you have learned to go through in the processes of multiplication or long division of numbers? It must be said that to some extent it was inevitable that you should learn to do many things in arithmetic, say, before you were old enough to understand them. But why not understand them now? Algebra is not merely a compendium of tricks for solving catchy problems. If it were, it would deserve to be called "a low form of cunning." On the contrary, algebra has the same logical structure as geometry or any other branch of mathematics as you will see later. Mathematical problems assuredly have at least as much "puzzle appeal" as more popular hobbies such as card games, cross-word puzzles and the like, and are at the same time more valuable. But mathematics has a far deeper significance than could possibly be possessed by a mere collection of puzzles.

In the early part of the course we shall discuss many things in algebra and geometry which you have already studied in your high school courses, but from a very different point of view. We shall seldom ask you to remember accurately any of the techniques you took up in the past and even arithmetic will be discussed here from a logical standpoint. In fact, in some parts of the book, you will find it necessary to forget many things which you have learned to do by habit in order to be able to consider their logical justification without preconceived prejudice. Later in the course you will be introduced to some of the easier fundamental ideas of many fascinating, modern, and so-called advanced branches of mathematics, although we shall never develop these subjects far enough to make stringent demands on your technical equipment. While, of course, there is no easy road to mastery of the mathematical sciences, many of the fundamental ideas, about which the intelligent layman is often curious, can be readily grasped without too much technique. We think that these things will prove more interesting to you and will be of immeasurably greater value to your mental develop-

ment and your appreciation and understanding of all the subjects forming part of the search for truth, than would, for example, the study of complicated equations.

While most people are somewhat aware of the importance of the technical applications of mathematics to the civilized world, many who do not intend to be technicians are inclined to doubt the value of the study of mathematics to themselves. This attitude may be due partly to the widespread failure to understand the nature of mathematics and to grasp the distinction between mathematics and its applications. At the end of this course, when you understand better the nature of both pure and applied mathematics, we hope you will be more inclined to agree with Benjamin Franklin who wrote,* "Whatever may have been imputed to some other studies under the notion of insignificancy and loss of time, yet these (mathematics) I believe, never caused repentance in any, except it was for their remissness in the prosecution of them."

2. Advice to the student. This book should not be used merely as a collection of homework exercises; it should be read. *You should read each section at least twice; once to get the general idea, and then carefully, checking on each detail. A third reading is advisable.* Conciseness and accuracy of statement are characteristic of mathematical writing. Therefore, read slowly and make an earnest effort to grasp what you read. *Reading with a pencil and paper in hand is strongly recommended; work out the illustrative examples in the book and make up similar illustrations for yourself.* Your chief burden will not be the task of working out large numbers of exercises each day, but rather will be the necessity for sustained attention and a sincere desire to understand. Approach the subject without the dislike you may have felt for it in high school. Its treatment will probably be so different that it will be, at times, difficult to recognize it as the same subject. Many people are fascinated by mathematics; it will be worth your while to give it a fresh chance.

Please do not approach even the most elementary parts of our work with the cock-sure sophisticated attitude that you "know this because you have already had it in high school"

* "On the Usefulness of Mathematics," *Complete Works of Benjamin Franklin*, vol. I, p. 421, G. P. Putnam's Sons, The Knickerbocker Press (1887).

and therefore do not have to read it carefully. Even the person who does not take the familiar radio, for example, for granted and who realizes that an understanding of radio depends on the mathematical work of many centuries, even such a person may yet take arithmetic for granted without realizing the long blundering struggle of the human race to evolve a suitable arithmetic. In this age of mechanical wonders, you must guard against subconsciously assuming that you understand something merely because you are familiar with it, or that things are true merely because you have habitually believed them. A wholesome, naive honesty is essential to the study of the fundamentals of science.

On the other hand do not approach the subject with an unreasonable fear. Although there is no lack of unsolved problems at the frontier of mathematical research, there are no mysteries in a completed chapter of mathematics. It is all simple, honest reasoning which anybody can follow, with a little effort and concentration, if it is explained step by step. Often a student has difficulty in following a mathematical explanation only because the author skips steps with which he consciously or subconsciously assumes the student to be familiar. If you do not recall the missing steps you may be faced with an unbridgeable gap in the reasoning and his conclusion may appear mysterious to you. We shall try to avoid such gaps here.

Another difficulty found by beginners in mathematics is the necessity for mastering the symbolism and technical vocabulary. This is to some extent unavoidable. The symbols are introduced to clarify and simplify statements and ideas which would seem very complicated if expressed in ordinary language. When you have learned to read the symbolism freely and fluently you can concentrate instead on the flow of ideas expressed therein. The situation is quite analogous to that of a beginner in music who cannot read and appreciate a great musical composition because he has to concentrate painfully on the meaning of individual notes, or a beginner in a foreign language who cannot appreciate the beauty of a great piece of writing because he has to concentrate painfully on the meaning of the individual words. Algebraic language, like any language, is merely a collection of symbols which we agree to interpret and combine in convenient ways. Therefore, to appreciate the great story of mathematical

thinking, *learn carefully the definitions of whatever symbols or unfamiliar terms we introduce*, and thereby leave your mind free to grasp the larger, fundamental ideas of the subject. Even where familiar words are used, their meaning in mathematical usage is often entirely different from their meaning in everyday usage. For example, no one who was unacquainted with the technical definitions could be expected to have the slightest conception of the mathematical meaning of the words "power," "root," "coordinate," "function," "group," "rational," "differentiate," "integrate," all of which will be used in this book. A person who persists in using words, which have been defined precisely, without understanding their definitions, literally does not know what he is talking about.

Keep in mind that it requires no unusual ability to follow mathematical ideas although it may have required genius to invent them. There is a great difference between breaking a trail through an uncharted wilderness and riding through it on a well-paved and graded road after it has been conquered. To create new mathematics of value often requires the highest order of imagination and intuition, not wild and untrammeled but subject to the final test of reason. (Thus new discoveries are not always the result of logical analysis alone but are often due to insight, vision, and the happy "accident" of hitting upon the appropriate path. However, it should be noted that these "accidents," as a great mathematician once remarked, seem to happen only to deserving people.) But to follow an explanation requires only close attention, the requisite background, and a normal ability to reason. In this book we shall demand little of your background. The rest is up to you. Mathematicians are no more than human; keep in mind the proverb: "What one fool can do, another can."

Chapter II

LOGIC, MATHEMATICS, AND SCIENCE

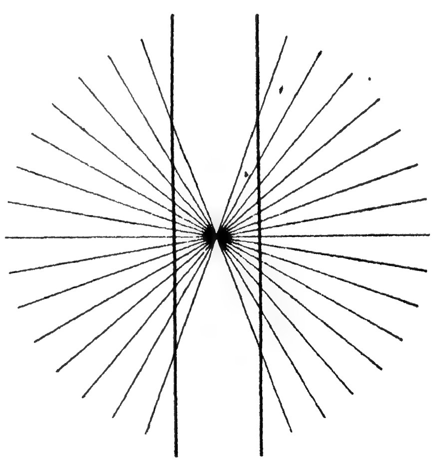
3. Introduction. You have recalled that the geometry of the ancient Greeks which you have studied in high school was not merely a collection of unconnected assertions but that all of them after the first few were deduced logically from those first few which, in turn, were assumed without proof. If one characteristic of mathematics is to be singled out as being of outstanding importance, it would certainly be the insistence on logical proof. Since logic plays so important a part, let us turn our attention to logic itself.

Logical thought is neither as common nor as easy as may be supposed. We all know that we believe many things without ever having satisfied ourselves logically about them. For example, the principal reason that most of us have for believing that the earth is roughly spherical is that people whose opinions we respect, like our scientists and teachers, tell us it is so.* But at one time, nearly everybody believed the earth to be flat, for the same inadequate reason. This state of affairs is largely inevitable simply because life is too short for us to examine critically everything we believe. Even so, it could do us no harm to be aware of what we are accepting uncritically on someone's authority. However, while most of us admit freely that we *do not* always think logically, many people are of the opinion that

* The reasons commonly advanced by students are hopelessly inadequate. For example, the gradual disappearance of ships at sea from hull to mast-top may be regarded as evidence that the earth is not flat (assuming that light travels in straight lines, etc.) but does not establish sphericity; this phenomenon is just as consistent with other shapes such as the shape of a football, an egg, etc. The same remark can be applied to the argument that the earth can be circumnavigated. The circular edge of the earth's shadow on the moon during eclipses is also inconclusive evidence since other shapes than spheres may well cast a shadow with a circular edge. The approximate sphericity of the earth is really established by extensive geodetic surveys showing that the "curvature" of its surface is approximately the same at all points. It follows from a theorem of higher mathematics that the only closed surface with constant positive curvature is a sphere.

they are *capable* of thinking logically quite automatically or instinctively and without effort, care, concentration, or study. This opinion is much more prevalent among freshmen than among mathematicians who have really attempted to think logically at least part of the time. Mathematicians are painfully aware of the effort required for the task even when the material is written in the comparatively transparent style of mathematical works where one's attention is deliberately directed to the openly exhibited logical structure of the argument. It must be apparent to a reflective person that it is much more difficult to reason logically in subjects where the reasoning, far from being openly exhibited in its skeletal form, is hidden in elaborate verbiage, and where our thoughts are colored by our emotions and prejudices.

The fact is that most of us, including professional mathematicians in their unguarded moments, seem to derive an almost athletic exhilaration from the popular sport of jumping to conclusions. There are even some extremists who claim that intuition is a good substitute for reason. Now intuition has an important rôle in science, especially as a "suggestor," but it certainly does not take the place of reason. The creative mathematician uses imagination and intuition to conjecture new results and new methods of research, but he does not assert that his guesses are correct until he has succeeded in proving them logically. Even a confirmed mystic would hesitate before driving his car over a bridge constructed exclusively by intuition. In fact, simple things like optical illusions teach us not to draw hasty conclusions from the evidence of our senses (Fig. 1).



Apply a ruler to the "bent" vertical lines

Fig. 1

The results of logical reasoning in the investigation of natural phenomena have amply justified the importance given it in mathe-

matics and science in general. None of the alternatives to logic (authority, intuition, or mysticism) has a record of achievement in these fields as enviable as that of logical reasoning. In the remainder of this chapter, we shall, therefore, examine more closely the nature and importance of logical reasoning and its connection with mathematics and science in general.

EXERCISES

Answer the following questions intuitively; then check your answers by calculation if you can.

1. If A can do a job in 6 days and B can do it in 10 days, how long will it take them to do it together?

2. A man sells 60 pieces of candy at 3 for 1 cent and another 60 pieces at 2 for 1 cent. Would his income be more, less, or the same if he sold 120 pieces at 5 for 2 cents?

3. A car travels 10 miles at 30 miles per hour and the next 10 miles at 60 miles per hour. Would the trip take more, less, or the same time if it travelled all 20 miles at the steady rate of 45 miles per hour?

4. Each bacterium in a culture splits into two bacteria once a minute. If there are 100 billion billion bacteria present at the end of an hour, when were there exactly 50 billion billion present?

5. Two civil service jobs have the same starting salary of $1800 per year and the same maximum salary of $3000 per year. One gets an annual raise of $200; the other gets a semi-annual raise of $50. Which is better?

6. A square sheet of copper one foot on each side costs 20 cents. What will be the cost of a square sheet of copper one and one half feet on each side?

4. Deductive logic. Truth and validity. We shall make no attempt here to discuss the philosophical questions "What is truth?" and "How can we determine whether or not a given statement is true?" In fact, it is exceedingly doubtful whether either question can be answered with anything like finality. We shall suppose that we can determine somehow, in some cases, whether given statements are true or false; or at least we will be willing to agree or grant (tentatively perhaps), or simply to assume, that some statements are true and others false. We shall be principally concerned with the problem of getting new truths from old. That is, granting that certain statements are true, what can be said about the truth of certain other statements? This is essentially the problem of logic.

If two statements are so related that the second *must* be true *if* the first is true, then we say that the second **follows from** the

first, or the second is a **logical consequence** of the first, or the first **implies** the second. The first statement is called the **hypothesis** and the second is called the **conclusion.**

Example 1.

Hypothesis: (*a*) All freshmen are undergraduates.
(*b*) All undergraduates are geniuses.

Conclusion: All freshmen are geniuses.

Note that the expression "All A's are B's" means that "every A is also a B." It does not mean that every B is also an A. If all A's are B's and all B's are A's then the class of A's is identical with the class of B's.

Here the hypothesis implies the conclusion, since it is clear that *if* the hypothesis is true *then* the conclusion *must* be. Notice that there may be some doubt about the truth of the conclusion and of part (*b*) of the hypothesis.

If the conclusion of an argument really follows from the hypothesis, the argument is called **valid.** The process of drawing valid conclusions from given hypotheses is called **deduction** or **deductive reasoning.** We must distinguish carefully between the actual *truth* of the statements involved and the correctness or *validity* of the argument. A good way to test the validity of an argument like the above example is to use diagrams like the following.

Represent the class of all freshmen by the set of all points inside a closed boundary line. Do likewise with the class of all undergraduates and the class of all geniuses. Part (*a*) of the hypothesis demands that the class of freshmen be entirely contained within the class of all undergraduates. Part (*b*) of the hypothesis demands that the class of undergraduates be entirely contained within the class of geniuses. Thus the various sets of points representing these classes of people must be placed in the relative positions pictured in Fig. 2. Clearly, the requirements of the hypothesis have *forced* us to place the class of freshmen entirely within the class of geniuses, which is exactly what is as-

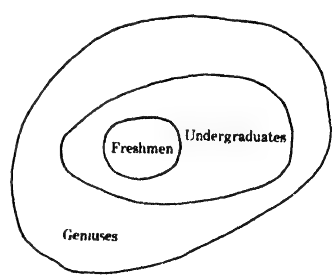


FIG. 2

serted in the conclusion. Hence the argument of example 1 is valid.

We say that the reasoning is valid regardless of the actual truth of the statements involved. Valid reasoning tells us that *if* the hypothesis *were* true, then the conclusion *would have to be* true. For this reason it is often called "*if* . . . *then* . . ." reasoning.

Example 2.

Hypothesis: (*a*) All freshmen are human.
(*b*) All undergraduates are human.

Conclusion: All freshmen are undergraduates.

There is no disagreement as to the truth of all these statements. But the reasoning is *not* valid, because the hypothesis does not *force* you to place the class of freshmen entirely within the class of undergraduates. You are forced *only* to place the class of freshmen within the class of humans, as in Fig. 3. Therefore the argument is not valid.

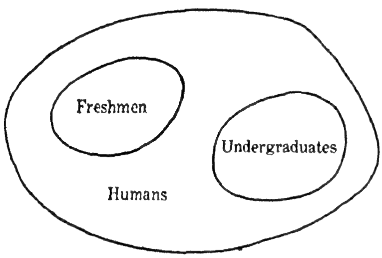


Fig. 3

Of course it *may* happen that all freshmen are undergraduates. But we are here concerned with the question whether it *must* happen because of our hypothesis. It is important to grasp the distinction between *may* and *must.*

If it is *possible* to satisfy the requirements of the hypothesis *without* satisfying the conclusion then the argument is *not valid.* Reasoning is called valid only if the conclusion is inescapable.

In valid reasoning the conclusion must be true if the hypothesis is true. Does this imply that the conclusion must be false if the hypothesis is false? Consider the following example.

Example 3.

Hypothesis: (*a*) All New Yorkers are Martians.
(*b*) All Martians are residents of the United States.

Conclusion: All New Yorkers are residents of the United States.

Here the reasoning is valid, the hypothesis is false, but the conclusion is true. See Fig. 4.

As another example, consider the proposition "if $3 = 7$ then $1 = 1$." Granting that if equals are divided by equals the results are equal, we can deduce the conclusion as follows:

$$\begin{aligned} \text{if } 3 &= 7 \\ \text{then } 3 &= 7 \\ \text{and } \frac{3}{3} &= \frac{7}{7} \\ \text{or } 1 &= 1. \end{aligned}$$

The reasoning is valid and the conclusion is true, but the hypothesis is false.

Thus a false hypothesis *may* yield a true conclusion. *The truth of the conclusion does not imply the truth of the hypothesis.* This is an important though elementary point to which we shall refer later. That is, *if statement A (the hypothesis) implies statement B (the conclusion) and B is known to be true, then we have no right to make any inference as to whether A is true or not;* that is, the truth of A is not guaranteed by the truth of B.

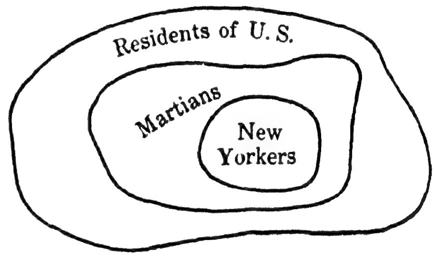


FIG. 4

The **converse** of the statement "A implies B" (or "if A then B") is the statement "B implies A" (or "if B then A"); that is, *the converse is formed by interchanging hypothesis and conclusion.* If we knew that "B implies A" was valid and B were known to be true then A would have to be true. But the converse of a valid argument may well fail to be valid, and the converse of a true statement may well be false. For example, a valid argument is the following: "if $x = 2$ and $y = 5$ then $x + y = 7$." The converse "if $x + y = 7$ then $x = 2$ and $y = 5$" is clearly not valid for we might have $x = 3$ and $y = 4$ or some other combination. "If an animal is a dog, it is a quadruped," is true; the converse, "if an animal is a quadruped, it is a dog" is false, as all the donkeys in the world attest. Only statements of the "if . . . then . . ." type have converses. "This apple is red" has no con-

verse. The statement "all a's are b's" has the converse "all b's are a's"; note that "all a's are b's" means that "if anything is an a then it is a b" and is therefore of the "if . . . then . . ." type.

Example 4.

Hypothesis: (a) No interesting people are bores.

(b) Some interesting people are college professors.

Conclusion: Some college professors are not bores.

This is a valid argument, regardless of anyone's opinion as to the truth of the statements involved.

Note that "no A's are B's" means that the class of A's and the class of B's have no members in common. The expression "some A's are B's" means "there exists at least one A which is also a B." Notice how these statements are pictured in Fig. 5. The statement "some A's are B's" does not imply that "some A's are not B's"; it is merely non-committal as to the existence of A's which are not B's. One must be careful not to be misled into reading such a conclusion from the figure. This meaning of "some" is different from its everyday usage where it means "some but not all." The non-committal usage is more useful for logic because it enables us to deal with situations in which we cannot be sure whether all A's are B's or not but where we are sure that at least one A is a B.

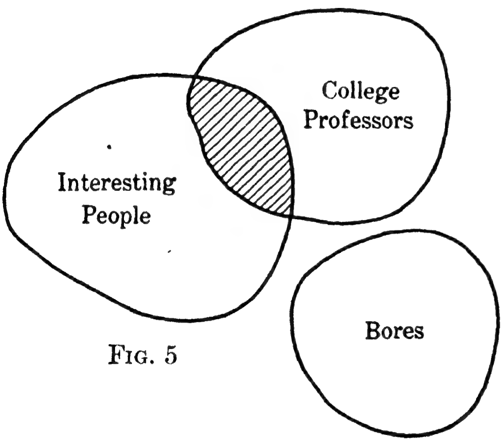


FIG. 5

A detailed study of the logic of arguments of the type of most of the examples in this section was made by Aristotle (4th Century B.C.). The diagrams used here are ascribed to the 18th Century Swiss mathematician L. Euler. These diagrams, while convenient, are not essential to the reasoning. They are merely used to picture the relationships among various classes of objects, such as the relationship "class A is included within class B." The study of logic progressed materially beyond the state in which

Aristotle left it only in modern times when mathematicians became interested in it.

EXERCISES

Test the validity of the following arguments by using diagrams and give your opinion as to the truth of the various statements involved.

1. *Hypothesis:* (*a*) All intelligent students can pass mathematics.
(*b*) Jones can pass mathematics.
Conclusion: Jones is an intelligent student.

2. *Hypothesis:* (*a*) Some college students are clever.
(*b*) All freshmen are college students.
Conclusion: Some freshmen are clever.

3. *Hypothesis:* (*a*) All college students are clever.
(*b*) All freshmen are clever.
Conclusion: All freshmen are college students.

4. *Hypothesis:* (*a*) All clever people are college students.
(*b*) All freshmen are clever people.
Conclusion: All freshmen are college students.

5. *Hypothesis:* (*a*) No conceited people have a sense of humor.
(*b*) Some teachers do not have a sense of humor.
Conclusion: Some teachers are conceited.

6. *Hypothesis:* (*a*) No college students are maniacs.
(*b*) All freshmen are college students.
Conclusion: No freshmen are maniacs.

7. *Hypothesis:* (*a*) All maniacs are college students.
(*b*) No freshmen are college students.
Conclusion: No freshmen are maniacs.

8. *Hypothesis:* (*a*) All mumbos are jumbos.
(*b*) All jumbos are boojums.
(*c*) All boojums are snarks.
Conclusion: All mumbos are snarks.

9. Test the validity of the following argument by using diagrams. Do not give your opinion as to the truth of the statements involved.

Hypothesis: (*a*) All mathematics instructors are absent-minded people.
(*b*) All absent-minded people are at least slightly crazy.
Conclusion: All mathematics instructors are at least slightly crazy.

10. Test the validity of *each* of the proposed conclusions by using diagrams.

Hypothesis: (*a*) All timid creatures are bunnies.
(*b*) Some timid creatures are dumb.
(*c*) Some freshmen are timid creatures.
Conclusion: (*a*) Some bunnies are dumb.
(*b*) Some freshmen are bunnies.
(*c*) Some freshmen are dumb bunnies.

11. Test the validity of *each* of the proposed conclusions by using diagrams:

Hypothesis: (*a*) All cats are bats.
(*b*) Some bats are gnats.
(*c*) Some bats are grey.

Conclusion: (*a*) Some cats are gnats. ✓
(*b*) Some gnats are grey. ✓
(*c*) Some cats are grey. ✓

12. (*i*) Test the validity of the following argument.

Hypothesis: (*a*) All murders are immoral acts.
(*b*) Some murders are justifiable.

Conclusion: Some immoral acts are justifiable.

(*ii*) If part (*a*) of the above hypothesis is true and the conclusion is false, what can be said about the truth or falsity of part (*b*) of the hypothesis? Explain.

13. Make up four true statements of the "if . . . then . . ." type, which have false converses.

14. Assuming that the conclusion follows from the hypothesis by valid reasoning, complete each of the following sentences with one of the phrases "must be true," "must be false," "may be true or false":

(*a*) If the hypothesis is true, then the conclusion must be true
(*b*) If the hypothesis is false, then the conclusion may be [illegible]
(*c*) If the conclusion is true, then the hypothesis may be [illegible]
(*d*) If the conclusion is false, then the hypothesis [illegible]

15. If statement A implies statement B, and statement B implies statement C, must A imply C? Explain. Yes [illegible]

16. If statement A implies statement B, must B imply A? Explain. NO

17. If a person reasons correctly and obtains a conclusion with which you agree are you logically forced to agree with his hypothesis? Explain.

18. Criticize the following argument: "If Brown had murdered Mrs. Grundy he would have acted exactly as he did when he met Jones at the club. Therefore Brown is the murderer." [illegible]

5. The formal character of deductive logic. We have seen that the validity of an argument does not depend on the actual truth of its statements. More than that, validity depends only on the *form* of the statements and not on their meaning. The statements of a valid argument may be false or they may have no meaning at all. For instance, the arguments of example 1, section 4, and of exercise 9, section 4, both have the following abstract form.

Example 1.

Hypothesis: (*a*) All x's are y's.
(*b*) All y's are z's.

Conclusion: All x's are z's.

This abstract form of argument is *valid* regardless of what concrete meanings you give to x, y, z. To speak of the truth of its statements is clearly absurd since its statements do not have any meaning until we assign to (or substitute for) the undefined terms x, y, z some definite meanings. Reasoning with such meaningless forms is called **formal logic** or **abstract logic.**

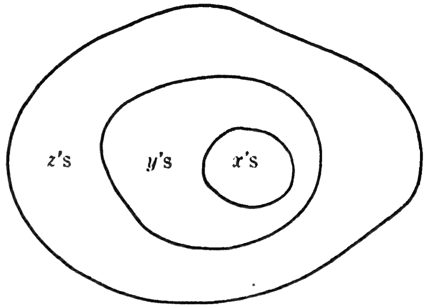


FIG. 6

Some philosophers have scoffed at this so-called "thoughtless thinking." But it is precisely the fact that abstract logic is independent of any particular subject matter that makes it so valuable. For it can be applied equally well to any subject matter. If you substitute for x, y, and z, in example 1, any meanings which convert the hypotheses into true statements, then you know automatically that the conclusion is also converted into a true statement. This applies equally well to a simple argument or to a long complicated chain of such arguments. In formal or abstract logic we have the additional practical advantage of being able to decide whether or not reasoning is valid without being disturbed or influenced psychologically or emotionally by the meaning of the statements, since they have no specific meaning and may be given many different meanings. It is much easier to tell good reasoning from bad when the argument is openly exhibited in its skeleton form than when it is surrounded by sonorous rhetoric and deceptive verbiage and when the meanings of the statements are apt to color our thoughts.

Remark 1. Statements like

(1) "All x's are y's,"

or

(2) "x was the first President of the United States,"

cannot be considered assertions at all until meanings have been assigned to the undefined terms x and y; when meanings *have* been assigned, the resulting statements may be true or false or merely nonsense. Thus, if we substitute x = "Abraham Lincoln" in (2) we get a false statement; if we substitute x =

"George Washington" in (2) we get a true statement; if we substitute x = "imperceptibility" in (2) we get nonsense. The distinction between a false statement and a nonsensical one is not always easy to make. A symbol like x, y, or z, which may take on more than one meaning is called a **variable.** Statements like (1) and (2) which contain undefined terms or variables are often called **propositional functions;** they become **propositions** only when we have substituted meanings for the undefined terms which make the resulting sentences either true or false. A nonsensical statement is not considered a proposition at all. Formal logic is really concerned with propositional functions and with the validity of arguments involving them. When we speak of truth or falsity we are referring to propositions. The sentences on an application form which contain blank spaces are purely formal; that is they are propositional functions in which the blanks are the variables. When we substitute meanings for these variables (fill in the blanks), these sentences become propositions.

Remark 2. It is important not to be deceived by the similarity of the statements:

(1) If A is true then B is true;

(2) If B is true then A is true;

(3) If A is not true then B is not true.

The unwary often confuse these three. *But* (1) *does not imply* (2), *and* (1) *does not imply* (3). *Likewise* (3) *does not imply* (1) *and* (2) *does not imply* (1). The statements (1) and (2) are *converses* of each other. Two statements A and B are called **equivalent** if A implies B *and* B implies A. Hence if (1) and (2), or (1) and (3), are both correct, then A and B are equivalent. Statement (1) is often written in the form "A is true **only if** B is true." Thus, the statement "A is true **if and only if** B is true" means that "A and B are equivalent"; it is a compact way of stating (1) and its converse (2) together.

Example. Consider the following statements. (1) If a person is reading this book then he is alive. (2) If a person is alive then he is reading this book. (3) If a person is not reading this book, then he is not alive.

Remark 3. Many arguments found in print are either absolutely incorrect, or, at best, can be made valid only by the addition of further hypotheses, not stated by the author. Such situations can often be detected by substituting symbols like x, y, and z for the terms in the argument and putting the argument into a form which exhibits its logical structure.

Example. Consider the following argument. "Brown has publicly asserted that he believes in democracy. Therefore he is a good citizen because all who believe in democracy are good citizens." Putting this argument into a form which exhibits its logical structure, we have:

Hypothesis: (*a*) Brown has publicly asserted that he believes in democracy.
(*b*) All who believe in democracy are good citizens.

Conclusion: Brown is a good citizen.

This is certainly *not* valid as it stands. To see this clearly let us introduce symbols as follows. Let B represent Brown; let P be the class of all those who have publicly asserted that they believe in democracy; let G be the class of all good citizens; and let D be the class of those who believe in democracy. Note that P and D are different classes. Our argument becomes (see Fig. 7):

Hypothesis: (*a*) B is in P.
(*b*) D is contained in G.

Conclusion: B is in G.

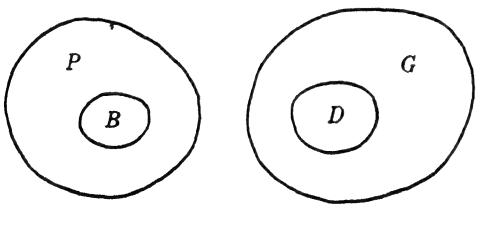


FIG. 7

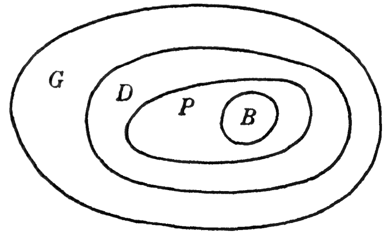


FIG. 8

However, this invalid argument will be changed into a valid one *if* we make the further hypothesis that all who publicly assert that they believe in democracy do believe in it; that is, that P is contained within D (Fig. 8).

EXERCISES

Test the validity of the following arguments.

1. *Hypothesis:* (*a*) All x's are z's.
(*b*) All y's are z's.
Conclusion: Some x's are y's.

2. *Hypothesis:* (*a*) Some x's are y's.
(*b*) All x's are z's.
Conclusion: Some z's are y's.

Put each of the following arguments into a form which will exhibit its logical structure, and decide whether or not it is valid as it stands. If not, see if you can make a valid argument out of it by assuming additional hypotheses not stated here.

3. All humane men are kind to domestic animals. Brown is humane because he is kind to domestic animals.

4. People who make mistakes cannot be trusted. It is silly to ask advice of people who cannot be trusted. All physicians make mistakes. Therefore it is silly to ask advice of a physician.

5. Helping the poor is always good. Therefore giving alms to beggars is always good.

6. Brown's program is unworthy of support. Jones' policies are worthy of support because he is unalterably opposed to Brown's program.

7. A classical education is worthless because we make no practical use of Latin and Greek after graduation.

8. This is an evil theory for we find it supported by some evil men.

9. People with inferiority complexes are always pompous. Professor Bluccester is always pompous. Therefore, he has an inferiority complex.

10. Assuming that the statement "If Smith is a college student, then he is more than 12 years old" is true, which of the following statements is necessarily true?

(*a*) If Smith is more than 12 years old, he is a college student.
(*b*) If Smith is not a college student, he is not more than 12 years old.
(*c*) If Smith is not more than 12 years old, he is not a college student.

11. (*a*) Make up an example in which statements (1), (2), and (3) of Remark 2 are all correct.

(*b*) Make up an example in which they are not all correct.

12. Find some examples of invalid arguments in your outside reading.

6. Inductive logic and experimental science. Inductive logic is the name given to the process of coming to a *probable* general conclusion on the basis of (many) particular instances. It is thus a tremendously important weapon of experimental or empirical science and is used unconsciously in everyday life. None

the less its conclusions are never certain but only more or less probable. For example, if a man performs repeatedly the simple experiment of throwing a pair of dice and he throws a total of seven five times in succession, we are only mildly surprised. But if he throws a total of seven 20 times in succession we form the tentative conclusion that these particular dice will always turn up with a total of seven; that is, they are loaded. However, this conclusion is not really certain, but only probable. It seems to become more probable if it happens 50 times in succession, so probable that many of us would act on the assumption of the truth of our conclusion. Nevertheless, it *may* not be true and we may merely have witnessed an unusual run of luck; there is nothing in the theory of probability that prevents it. We shall return to this topic in Chapter XIII. Despite the importance of inductive logic in experimental science we shall refer to it seldom. Of supreme importance for mathematics is deductive logic with which we shall be almost exclusively concerned hereafter.

Remark. A **general statement** is one referring to *all* the members of some class of things; a **particular statement** refers only to *some* of the members of some class of things. It is commonly said that deductive reasoning obtains particular statements from universal or general statements while inductive reasoning obtains general statements from particular statements. Neither of these descriptions is appropriate, according to our definition of deductive reasoning. In deduction, one's conclusion may be general or particular. For example, to deduce the particular conclusion "Socrates is mortal" from the general hypothesis "All men are mortal" one must have the additional particular hypothesis "Socrates is a man." Similarly, consider the so-called inductive argument in which we seem to get a general conclusion from a particular hypothesis:

Hypothesis: (*a*) In the 500 cases we tried, treatment T cured disease D.

Conclusion: Treatment T will always cure disease D.

This can be made valid by adding the general hypothesis (*b*) that whatever cured our 500 cases will be a cure for all cases of disease D. With this additional hypothesis, the argument becomes deductive and the conclusion is inescapable. Now this

additional hypothesis is seldom asserted because it is seldom believed to be true. In fact, what many people refer to as induction is the assertion of this additional hypothesis. But if we are to distinguish between the logical question of whether our conclusion really follows from our hypothesis (a) and the psychological question of what is likely to convince the average mind, then we must emphasize that our conclusion is not valid unless we know that our 500 cases are really typical of all cases. Since this is seldom known, we usually say that our conclusion is more or less probable. What this means will be discussed further in Chapter XIII. More complete discussions of these considerations will be found in the references at the end of the present chapter.

Instead of trying to distinguish between deductive and inductive logic according to whether the conclusion is general or particular, it is better to distinguish between them according to whether the conclusion is necessary or merely probable on the basis of the hypothesis. *The conclusion of an inductive argument is never more than probable. The conclusion of a deductive argument is inescapable if the hypothesis is granted.*

7. Geometry. The first interest in geometry probably arose because of the social need for measurement of tracts of real estate, the problems of architecture, and the agricultural need for a calendar, which involves a knowledge of astronomy. We do not know with certainty how far back the beginnings of geometry go, but the Babylonians and Egyptians, from perhaps 3000 B.C. to about 500 B.C., certainly possessed some geometric knowledge and a remarkably good calendar. It is sometimes said that the river Nile is responsible for the beginnings of geometry. Certainly, the Nile overflowed periodically, and changed its shoreline. The Greek historian Herodotus (5th Century B.C.) conjectured that geometry originated because taxes on real estate in Egypt were paid in proportion to area, and, when the Nile's floods altered the shoreline, the neighboring estates had to be surveyed anew. In any case, while the Egyptian and Babylonian civilizations, long before the Greeks, had some information about geometry, they probably did not have the idea of logical proof. Their geometric theorems were largely uncon-

nected assertions, each established by empirical observation, at least approximately.* During the 7th and 6th centuries B.C. commercial intercourse sprang up between Egypt and Greece, and with it came an interchange of ideas. Greek scholars, visiting Egypt, learned what the Egyptians knew and made tremendous improvements. The statements of geometry were not regarded by the Greeks as unconnected propositions; rather, their purpose was to derive complex statements from simple ones and ultimately to derive all of them logically from the first few simple ones.

How did this important idea come into being? Let us try to blame this on the Nile as well. Imagine that the flooded Nile has erased the boundary lines between neighboring farms, and that Brown and Jones, two Egyptian farmers, who, since human nature was probably not very different then, are arguing about the boundary line between the farms. There is a rock at A on the river bank and the boundary is supposed to be at right angles to the bank at A. Brown claims AB is the boundary and Jones holds out for AC. How are they to resolve the dispute? A first method would be for one to kill the other and occupy both farms. This method, while fashionable among great nations, does not seem fair when applied to individuals. A second method would be to apply to the King or the High Priest for an arbitrary ruling. (This method resembles an attempt once said to have been made in the legislature of one of the United States of America to make the value of π exactly 3 in that state.) Both of these methods doubtless enjoyed great prestige in the past and to a slightly

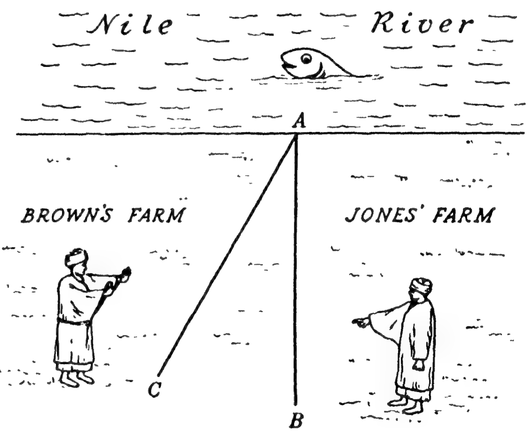


FIG. 9

* Some of their formulas were erroneous. Much of our knowledge of Egyptian mathematics comes from the Rhind papyrus written by Ahmes some time before 1700 B.C. and having the somewhat ambitious title: "Directions for obtaining the Knowledge of all Dark Things."

lower degree still do today, but they do not satisfy a reflective person. Therefore, Brown and Jones try to come to some common agreement as to where the perpendicular goes.

Jones' argument runs something like this: "The boundary is AC because I well remember that in the middle of that lake of mud, exactly at C, there used to be a bush beside which my favorite dog used to sleep, and I am ten years older than you, and, even if you did go to school in Alexandria, I've never yet admitted that anyone's judgment was as good as mine, and I'm not going to begin at my time of life, by Isis!"

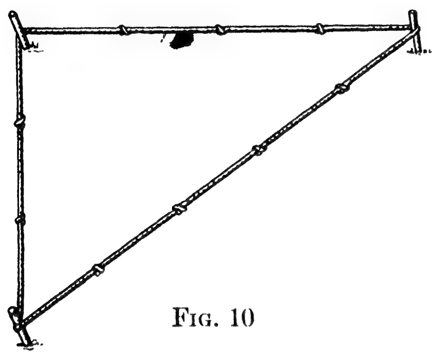


Fig. 10

Brown's argument is that if you take a rope with knots at equal intervals, and stretch it taut about three stakes so that the opposite sides contain 3, 4, and 5 intervals respectively, then the angle opposite the longest side has to be a right angle (this was probably known to the Egyptian surveyors, who were called "rope-stretchers"); and he had done this and found the boundary to be AB.

The following dialogue then takes place.

Jones: I don't believe that 3–4–5 hocus-pocus.

Brown: Do you believe that *if the lengths x, y, and z of the three sides of a triangle satisfy the equation $x^2 + y^2 = z^2$ then the angle opposite side z is a right angle?**

Jones: If that were true, your 3, 4, 5 rope-trick would be correct, since $3^2 + 4^2 = 5^2$, and you would win the argument; so I don't believe it.

Brown: Do you believe (*the Pythagorean theorem*) ** that *if*

* This is the converse of the Pythagorean theorem which you learned in school. You probably never proved this one yourself, but it can be deduced from the Pythagorean theorem as follows. Construct a right triangle the lengths of whose legs are x and y. Then its hypotenuse is equal to z by the Pythagorean theorem. Hence the second triangle is congruent to the given one since 3 sides of one are equal respectively to 3 sides of the other. Therefore the given triangle is also a right triangle.

** The Pythagorean theorem should be memorized. It is one of the most important theorems in elementary geometry.

x, y, z are the lengths of the sides of a right triangle, z being opposite the right angle, then $x^2 + y^2 = z^2$?

Jones: That sounds too much like the previous statement. I don't believe it.

Brown: Do you believe that *if two triangles are similar* * *then their sides are in proportion?*

Jones: If it gives you that land, I don't believe it!

Brown, seeing what he has to contend with, takes a breath and begins again.

Brown: Do you believe that through two points, one and only one line can be drawn?

Jones: Certainly, but what has that to do with the case?

Brown: Do you believe that if equals are added to equals, the results are equal?

Jones: Of course. But keep to the subject of this boundary—I'm a busy man.

However, Brown continues asking Jones whether he believes certain simple statements which happen to be axioms or postulates on page 1 of the geometry book which Brown had studied. Jones, impatient, and still failing to see any way in which those statements can cause him to lose the argument, admits that he believes them all. Thereupon Brown proceeds to prove, step by step, on the basis of these axioms, and with inexorable logic, that the converse of the Pythagorean theorem is valid and therefore that the rope stretching trick is legitimate, and the disputed land is his.

Now, this tall story is not accurate history, but it illustrates how the notion of proving a large body of statements on the basis of a few admitted statements might have arisen from the necessity for common agreement in very mundane, practical problems. On the other hand it might be due chiefly to the unusually logical bent of the Greek philosophers. This tremendously important idea of deducing large numbers of statements from a few simple assumptions probably grew slowly in the consciousness of the human race; it seems unlikely that it could have sprung full-grown from the brain of one man, although Thales (7th–6th

* Two triangles are called *similar* if the angles of one are respectively equal to the angles of the other.

centuries B.C.), one of the first of the great Greek mathematicians, is sometimes given credit for it. The earliest great written example extant of such extensively organized deductive thought is Euclid's *Elements* (about 300 B.C.), the work on which your high school geometry text was based. Euclid's *Elements* has long been regarded as representative of one of the outstanding achievements of the human mind, Greek geometry; but its full significance was probably not grasped until the 19th century. Euclid's chief contribution was not the discovery of new theorems but consisted in showing that all known theorems were logical consequences of a few assumptions. *)

Several practical advantages are derived from the deductive organization of geometry. For example, one might stare at right triangles for ages without perceiving the relation $x^2 + y^2 = z^2$ of the Pythagorean theorem; it is hardly self-evident. But having deduced it from simple assumptions one sees without effort that if the assumptions are true, so is the Pythagorean theorem. Thus instead of having to question the truth of hundreds of theorems separately, the question of their truth is made to depend on the truth of a few simple assumptions. In the other direction, one can deduce from theorems already in existence, new unsuspected theorems which might never have been found out by experimental observation. It is worth noting that Greek geometry, which aimed at understanding the logical interdependence of its statements, proved to be far more useful than, say, Egyptian geometry which aimed at being useful and nothing more.

8. Pure and applied mathematics. In this section we shall discuss two of the most basic terms in the entire book, namely *abstract mathematical science* and *concrete interpretation* or *application.* In terms of these two concepts we shall define *Pure Mathematics* and *Applied Mathematics.*

Notice that when logical structure is introduced into geometry, it is necessary to allow some of the statements to remain unproved (that is, not deduced from other statements), in order to have something to start with. Thus theorem 20 may be proved by showing that it follows logically from theorem 19; you then know that theorem 20 is true if theorem 19 is true. Similarly you may deduce theorem 19 from theorem 18, and so on. Sooner

or later you must stop and take some statements without proof, because we do not live forever. If you try to prove all your statements, you will necessarily commit the unpardonable sin of "circularity." That is, if you say that statement A is true because it follows from B and B is true, and then you assert that B is true because it follows from A and A is true, you are using circular reasoning. Putting your two steps together, you have said in effect that A is true because A is true. This is the mode of reasoning used by children who stamp their feet and shriek "It's so because it's so." If the truth of A is to be established by showing that A follows from B, then we cannot also establish the truth of B by showing that it follows from A. Hence, the entire structure of geometry must rest on some *unproved statements* called **axioms, assumptions,** or **postulates.**

Example. A king issues a proclamation saying that he is infallible. When asked to justify the truth of this statement, he answers that whatever he says must be true since the proclamation says that he is infallible. This reasoning is circular.

Now, the statements of geometry involve certain terms, such as "triangle." If we ask what this term means, we might define "triangle" in terms of other terms; thus we might say that a "triangle" is the set of all points lying on the line-segments joining three points which are not on the same (straight) line. This defines "triangle" in terms of the terms "point," "line-segment," etc. If we ask what these other terms mean, we might define them, in turn, in terms of still other terms; thus we might say that a "line-segment" is the set of all points lying between two given points. This defines "line-segment" in terms of "point," "between," etc. Sooner or later, this process of definition, like the process of proof, must stop. Clearly, we cannot define *all* our terms unless we fall into the trap of "circularity." We must use some terms without definitions, while all other terms may be defined ultimately in terms of these **undefined terms.** That is, all our definitions must involve some terms which are taken without definition.

Remark 1. The dictionary apparently defines all words. But it must therefore use circular definitions. For example, if you look up the word "insane," it might say that "insane" means

"crazy"; but if you look up "crazy," you might find that "crazy" is defined to mean "insane." Together these definitions inform you that "insane" means "insane." This situation is inevitable, as we have seen above, if you try to define all your terms.

Example. To define what we mean by saying that two rods "have the same length," we might say that this is so if one can be moved so as to coincide with the other. But if they were made of rubber, they could be made to coincide in any case. Hence we might go further and say that the two rods have the same length if one can be "moved rigidly" so as to coincide with the other. But how shall we define "moved rigidly"? We usually say that a rod is "moved rigidly" if it is moved in such a way that its "length" remains unchanged. We are saying, in effect, that two rods have the same length if one can be moved so as to coincide with the other in such a way that it has the same length at all times during the motion. This is clearly a circular definition.

Remark 2. A definition, in mathematical usage, is simply an agreement to regard one expression (symbol or set of symbols such as a word or a phrase) as being *equivalent* to another (usually more complicated) expression. The expression defined can always be substituted for its definition, and vice versa, in any context whatever without introducing any distortion or vagueness of meaning (although it may well introduce grammatical awkwardness). A definition, in this sense, should be distinguished from a (partial) description. Many dictionary "definitions" are merely descriptions. For example, we have *defined* a triangle above. But the statement, "a triangle is a figure extensively studied by mathematicians" is a *description.*

Bertrand Russell once * described mathematics as "the subject in which we never know what we are talking about nor whether what we are saying is true." Many students heartily agree with this, in a mood of personal confession. But it is correct literally. We never know what we are talking about since all our definitions rest ultimately on some undefined terms; we never know whether what we are saying is true because all our proofs rest ultimately on some meaningless unproved statements

* *International Monthly,* Vol. 4 (1901), p. 84.

or assumptions involving these undefined terms. Moreover, this situation must come about if you take any subject and try to give it a logical structure. Suppose you assert any proposition in economics, say, and you are asked to prove it. You may prove it by deducing it from other propositions. If these are challenged in turn, you may prove them on the basis of still other statements. Ultimately, however, you are forced to stop and base the whole thing on some unproved propositions or postulates. Similarly if you are asked what a word means, you will ultimately be driven to rest all your definitions on some undefined terms. When you have done both these things you have made an *abstract mathematical science*, or *deductive science*, or *abstract logical system*, of your subject. That is, an **abstract mathematical science** is a collection of statements beginning with some unproved statements or postulates (hypotheses) involving some undefined terms (basic terms), in which all further statements follow logically from the postulates and all new terms are defined in terms of the undefined ones. Reasoning from postulates in this way is sometimes called **postulational thinking.**

You may ask, "How can we reason about undefined terms? What do we know about them if they have no meaning?" The answer is simple and important. *We know about these undefined terms exactly what we have assumed about them in our postulates, or unproved statements, no more and no less.* In fact we must be careful not to use subconsciously anything that we have not explicitly stated in our postulates. That we can reason about such things is not strange at all. We have already done it in example 1, section 5, which may be considered as a miniature abstract mathematical science, as follows.

Example 1. Let the undefined terms be x, y, z. Let us assume two postulates.

POSTULATE 1. All x's are y's.

POSTULATE 2. All y's are z's.

THEOREM 1. All x's are z's.

We have here an abstract mathematical science with 3 undefined terms, 2 postulates, and one theorem.

Example 2. Let the undefined terms be x, y, z, and w.

POSTULATE 1. All x's are y's.

POSTULATE 2. Some x's are z's.

POSTULATE 3. All y's are w's.

THEOREM 1. Some y's are z's.

DEFINITION 1. Any y which is also a z will be called a "v".

THEOREM 2. Some w's are v's.

This is an abstract mathematical science with four undefined terms, three postulates, one definition, and two theorems. It is decidedly a miniature abstract mathematical science but it exhibits all the characteristics of an abstract mathematical science. It begins with undefined terms and unproved postulates; but thereafter, all new statements are logical consequences of previous statements and all new terms, like the term "v," are defined in terms of previously used terms.

In plane geometry "point," "line," and others, are often taken as undefined terms and statements like the following are often taken as unproved statements or postulates.

POSTULATE 1. Given any two distinct points, there is at least one line containing them.

POSTULATE 2. Given any two distinct points, there is at most one line containing them.

These two postulates are usually stated together, in high school texts on geometry as follows: given two distinct points there is exactly one line containing them, or, two points determine a line.

DEFINITION. If two lines in the same plane have no point in common they are called **parallel.**

POSTULATE 3. Given a line l and a point P not on l, there exists one and only one line l' containing P and parallel to l.

This is called the *Euclidean parallel postulate.**

Many other postulates and undefined terms, which we shall not list here, are taken at the start of geometry. On the basis of these postulates we can deduce all the theorems of geometry by pure deductive logic.

Geometry consists of the logical consequences of a set of un-

* Postulate 3 was not literally the parallel postulate made by Euclid but is known to be equivalent to Euclid's postulate.

proved statements involving certain undefined terms. Since we are using "point" and "line" as undefined terms, we might have changed the words "point" and "line" to "mumbo" and "jumbo" respectively, and written postulate 1 as: "given any two distinct mumbos, there is at least one jumbo containing them." Or we might have written "point" and "line" as x and y respectively; then postulate 1 would become "given any two distinct x's, there is at least one y containing them." This cannot affect the validity of the reasoning in our proofs. Geometry, like any other abstract mathematical science, is a tremendous example of abstract reasoning, where we reason about undefined terms on the basis of unproved statements. The postulates now constitute our hypothesis and all the theorems constitute our conclusion.

You may fairly ask, "In what sense can this meaningless game, this thoughtless thinking, this mumbo-jumbo, be true? How can it have any bearing on the real world?" The answer, honest and simple, is that such an abstract mathematical science can never tell you that your conclusions are *true* but asserts only that certain logical arguments are *valid.* It says merely that the theorems are logical consequences of the postulates or assumptions. Every mathematical statement is of the form "A implies B." This means that *if* A is true, *then* B must also be true. If we can find *meanings* or *concrete interpretations* for the undefined terms which will make the assumptions become *true* when these meanings are substituted for the undefined terms, then all the rest of the theorems must automatically become true statements about these meanings. For example, if we could find objects in the real world which have the properties demanded of "points" and "lines" (or mumbos and jumbos) in our postulates for geometry, we would then know that all our theorems are true statements about these objects.

When meanings are given to the undefined terms of an abstract mathematical science, we have a **concrete interpretation** or **application** of the abstract mathematical science. The meanings assigned to the undefined terms may be understood intuitively or by means of partial descriptions, etc. If the meanings assigned to the undefined terms are such that the postulates become true statements when these meanings are substituted for the undefined terms, then all the theorems become true state-

ments about these meanings, automatically, since the theorems are logical consequences of the postulates. The abstract mathematical science remains valid, however, whether such a concrete interpretation can be found for it or not.

Example 3. In example 1, page 27, we may convert our abstract science into a concrete interpretation by assigning the following concrete meanings to our undefined terms: let x = man, y = mortal, z = fool.

Remark 3. An abstract mathematical science is a body of logically connected propositional functions, while a concrete interpretation is a body of logically connected propositions (see remark 1 of section 5). We may speak of the validity of an abstract mathematical science but not of its truth. The question of truth arises only in connection with concrete interpretations or applications.

Exercise. Can there be different concrete interpretations or applications of the same abstract mathematical science? Illustrate using (a) example 1, page 27; (b) example 2, page 27.

The totality of all abstract mathematical sciences is called **Pure Mathematics.** The totality of their concrete interpretations or applications is called **Applied Mathematics. Mathematics** comprises both Pure and Applied Mathematics.

We must distinguish,* for example, between geometry as a branch of pure mathematics and geometry as a branch of applied mathematics, applied to the "real world." It is in this latter sense that you have probably thought of geometry up to now. You have thought of points and lines, for example, as being the dots and streaks you draw on paper, and you may have thought of the axioms of geometry as self-evident truths. In geometry, as an abstract mathematical science, point and line are undefined terms. If we think of these terms as having the concrete meanings "dot" and "streak," we are dealing with a concrete interpretation of this abstract mathematical science. With this interpretation, our postulates are certainly not satisfied except approximately. For no matter how small you make the dots or how thin you make the streaks, they are very large

* This distinction was not clearly grasped by the ancient Greeks who seem to have perceived the need for unproved postulates but not for undefined terms.

smudges viewed under a magnifying glass and, for example, many streaks can be drawn through two dots, contrary to postulate 2 (Fig. 11). It is only in an approximate sense that you can imagine that the postulates seem to be satisfied. Of course, for practical work, a good approximation is all that is needed because no measurement can ever be exact, since our senses and measuring instruments yield only approximate measurements. This concrete interpretation does seem to work well for our practical sciences and as long as it does we may make use of it. However, as pure mathematicians, we do not say that our postulates are self-evident truths (although this was said until the 19th century). We say only that if the postulates *were* satisfied, that is, converted into true statements, by any interpretation of the undefined terms, the theorems *would be true* too.

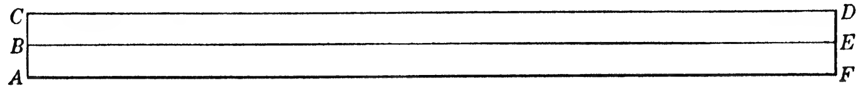
FIG. 11

In any concrete interpretation of an abstract mathematical science, we hope that the assumptions are true. We may indeed verify their truth approximately by observation, or we may feel that they are probably true, for one reason or another, but certainty concerning their absolute truth is often an unattainable ideal so far as science is concerned.

As for the view that our postulates are self-evident truths one may well ask, "Self-evident to whom?" It is notorious that what seems self-evidently true to one person may seem doubtful or false to another. If one says that we mean self-evident to a sufficiently large jury of competent people, we have only to reply that history is strewn with abandoned ideas which were once considered self-evident for intuitive or other reasons. For example: "the earth is flat"; "the heavenly bodies travel around the earth"; "the orbits of the planets around the sun are circles"; "every surface has two sides" (Fig. 12, p. 32); "a heavier body will fall faster than a lighter one." This last statement was asserted by Aristotle (4th century B.C.) and therefore believed by everyone until, according to some historians, Galileo (1564–1642) cast doubt upon it by actually dropping two bodies from a tower. They fell simultaneously.

In fact, even in ancient times *one* of the postulates of geometry (our postulate 3, the so-called Euclidean parallel postulate) was

not considered as self-evident as the others. This fortunate circumstance led to the invention of non-Euclidean geometries (in the 19th century) whose study contributed greatly to the understanding of the nature of pure mathematics and its applica-



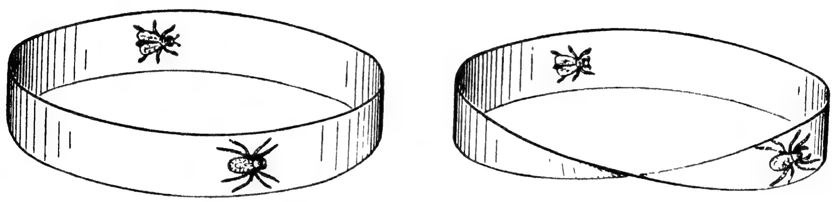
(a)

TWO-SIDED SURFACE

The spider cannot reach the fly without crossing the boundary. The surface has two sides and two edges.

ONE-SIDED SURFACE

The fly is not safe on this surface, which has only one side and one boundary edge.



(b) (c)

The surface (b) is made from the rectangular strip (a) of paper $ACDF$ by matching A with F, B with E, and C with D, and pasting. The surface (c) is made by twisting one end of the strip (a) through one half-revolution (180°) and matching A with D, B with E, and C with F, and pasting. The line BE is to be midway between CD and AF.

After the surface (c) has been made, try cutting it along the middle line BE. Then cut the resulting surface again along a line midway between *its* edges. Anyone who can guess what the resulting surface will look like has a far better intuition than most. Try it again, from the beginning, this time cutting along a line 1/3 of the width away from the edge.

FIG. 12

tion to the real world. We shall return to this subject in Chapter XVI.

The notion of mathematics as the totality of all abstract mathematical sciences and their concrete interpretations or applications is a tremendously broad conception, far transcending the old definition of mathematics as the science of space and quantity. In fact, it includes all subjects in which we reason logically, and amply justifies the assertion that mathematics is basic to all sciences. You may be reluctant to accept this strange new meaning of the word "mathematics." In fact, mathematicians accepted it with reluctance only toward the end of the

19th century when the development of modern mathematics forced this view on them.

We have seen that to construct an abstract mathematical science you have only to take a few undefined terms, assume a few compatible statements involving them, perhaps also define other terms in terms of them, and deduce all the logical consequences you can. But what undefined terms and assumptions shall we choose to start with, and, having chosen them, to which of the many possible logical consequences shall we give our attention, and what new terms shall we define? And even if we do decide all this, how, you may ask, can grown men waste their time on such empty games? You are partly right. Mathematicians are not much interested in an abstract mathematical science, except perhaps as a mental exercise, unless it shows promise of having interesting concrete interpretations or applications. By and large, we choose to work with certain postulates, definitions, etc., instead of others, because of the concrete interpretations or applications we have in mind for them. *The assumptions and terms we choose to work with and the theorems we try to prove are usually suggested by experience.* For example, they might be probable conclusions obtained by induction or they might seem to be self-evident facts. Or, their invention may be due to creative imagination, insight, or intuition. However, censorship should not be too strict. It would be a great mistake to restrict scientific research to those things for which we can find an immediate use. History has shown us many times that pure mathematics, worked out originally for its own sake, has found unsuspected applications 50 to 100 years later. For example, Riemann's ideas on geometry, developed early in the 19th century for their own sake, were applied to the physical theory of relativity by Einstein in the 20th century. Similarly, group theory, a branch of algebra developed largely in the 19th century with no thought of physical applications, is now used extensively in Quantum Physics. To cite an extreme case, certain curves, called conic sections, were studied by the ancient Greeks with no thought of application; they became of central importance in physical science about 1800 years later.

To summarize, an abstract mathematical science is constructed by selecting some undefined terms, some unproved

postulates or assumptions involving these terms, and then defining new terms and proving new statements or theorems. If we give concrete meanings to the undefined terms, we obtain a concrete interpretation or application of the abstract science. If we knew that, for these meanings, the postulates become true statements then we would know automatically that the theorems must also be true for this interpretation. But usually we are uncertain as to the truth of our postulates when they are interpreted concretely. For example, the notion that the postulates of geometry, interpreted concretely, are self-evident truths has been seen to be untenable. Reasoning from postulates is called postulational thinking. The totality of all abstract mathemathical sciences is called Pure Mathematics, and the totality of all concrete interpretations is called Applied Mathematics. Together they constitute Mathematics. We have seen that, in this broad sense, Mathematics includes every subject in which you try to be logical. For if you wish to prove a statement in any subject, you prove it by deducing it from other statements, and these in turn are either assumed or deduced from still other statements, and so on. Ultimately, if we are to avoid circular reasoning, the entire subject must be rested on some unproved statements or postulates. Similarly if we define a word it is defined in terms of other terms, and so on. Ultimately all our terms are expressible in terms of some which are left undefined, unless we fall into the trap of circular definition. When all this is done our subject has become a part of Mathematics, according to our definition of Mathematics.

We shall return to the ideas of abstract mathematical science and concrete interpretation many times in the remainder of the book and shall see many illustrations of these ideas, especially in Chapter XVII. See also section 19 for a non-trivial illustration.

9. Postulational thinking and scientific theories. We can now touch briefly on one aspect of the nature of scientific theories which often puzzles the layman. For example, how does it happen that Newton's theory of gravitation is discarded after being regarded as "true" for over 200 years? Newton's theory of gravitation can be regarded as an abstract logical system or abstract mathematical science, one of whose postulates, for ex-

ample, is the so-called first law * of motion: "Every body continues in its state of rest or of uniform motion in a straight line unless it is compelled to change that state by a force impressed upon it." From these postulates one may deduce logically a number of theorems which, when the undefined terms are interpreted in an obvious way, state how baseballs, bombs, planets, and pendulums should move. It may then be possible to verify some of these theorems (approximately) by empirical or experimental observation within the limits of reasonable error. But the postulates themselves usually cannot be verified by direct observation and are thus not self-evident.† For instance, heavenly bodies which are really at rest or in uniform motion in a straight line with respect to each other are entirely outside our experience. The chair on which you are sitting, for example, is undergoing a complicated motion, relative to the sun, resulting from the rotation of the earth on its axis, the revolution of the earth about the sun, etc. Hence we have the logical situation "A implies B and B is true," A being the postulates and B the experimentally verified theorems. But as we saw in section 4, this does not mean that A is really true.

In fact, we might well get the same theorems from different assumptions. Thus it should not surprise us to find two or more theories which "explain" observed phenomena equally well; for example, there were for some time two rival theories of light, the corpuscular and the wave theories. Nobody had ever seen a light corpuscle or a light wave directly. But assuming that one or the other of these things exists and has certain properties, one can deduce the experimentally verifiable behavior of light. As long as both systems continued to fit the facts equally well, one theory was as good an "explanation" as the other. However, should an observable phenomenon occur which contradicts one of our deduced theorems, we then have the logical situation A implies B and B is false; hence A must be false. Therefore, we must abandon or modify our abstract system by changing the postulates. If we do this well and arrive at a new abstract

* The word "law" is often used indiscriminately, in scientific writing, to refer to a postulate, theorem, definition, or a probable truth obtained experimentally.

† Even if they could be regarded as approximately verifiable truths for small billiard balls on a table, it would still be quite a jump to assume that they therefore hold for planets, etc.

logical system which fits not only the old observed facts but the new troublesome facts as well, then we say that we have a better theory than before. This is what happened at the beginning of the 20th century, for example, when observed facts turned up which did not fit Newton's theory of gravitation. The orbit of the planet Mercury * did not behave as predicted by Newton's theory; and the discrepancy was greater than could be attributed to observational error. There was little use in attempting to persuade the planet Mercury to get back to its course. Although Newton's theory had fitted all the facts for over 200 years, there was nothing to do but change it since we cannot change or ignore the facts. This was done in a remarkable way by Einstein whose new theory fitted the new facts as well as the old.† When and if new facts turn up which do not fit into Einstein's theory, it too will have to be abandoned or changed. Similarly "models" of the interior of an atom are hypotheses which are abandoned as soon as their logical consequences no longer fit the facts.

A striking explanation of the postulational nature of scientific theories is given in the following quotation from *The Evolution of Physics* by A. Einstein and L. Infeld: "In our endeavor to understand reality we are somewhat like a man trying to understand the mechanism of a closed watch. He sees the face and the moving hands, even hears its ticking, but he has no way of opening the case. If he is ingenious he may form some picture of a mechanism which could be responsible for all the things he observes, but he may never be quite sure his picture is the only one which could explain his observations."

But neither the question of whether the mechanism, which he has supposed hypothetically to be inside the case, is *really* there, nor even the interesting philosophical question of whether he may be sure that there is anything at all inside the case, need trouble the practical scientist as long as he has a theory which enables him to predict successfully the phenomena which can be observed.

Remark. We have said that our hypotheses or postulates are

* Among other things.

† The discrepancies between Newton's theory and Einstein's are too small to be observed except in connection with astronomical distances. Hence for the local purposes of engineering we still prefer to use the simpler theory of Newton.

usually suggested by observation. Some people are fond of saying that observed facts are *all* that matter and they use the word "theory" in a derogatory sense, saying, "Oh, that's only a theory." It is worth remarking that observation or empirical science cannot be wholly free from "theoretical" hypotheses. Experiment is always guided by theory. For even in collecting facts, we must have some hypothesis as to which facts are relevant to the investigation in hand, since we can hardly amass all the facts in the universe. In fact, we often do not trust our direct observations but rather correct them because of theoretical considerations.

We shall return to these ideas again when we have more mathematical background. Let us emphasize once more that postulational thinking is available for and indeed essential to a completely logical study of any subject matter. However, not many books, which pretend to be logical, outside of the so-called mathematical sciences, will bear critical examination from the point of view of postulational thinking. For instance, in reading a book on some social question, it would be a great comfort to know what underlying assumptions the author wants you to take for granted. There must be some, if he intends to prove anything logically. But he seldom tells you the basis for his "reasoning." In fact, if you try to imagine what his assumptions might be, in order to justify his conclusions, you are likely to find that one set of assumptions is needed on one page and an entirely different set on the next. Likewise, the meanings of his terms are likely to change from page to page.

Of course when we say that any logically organized subject "begins" with postulates and undefined terms, we mean "begins" in the logical, not the chronological, sense. Historically, every science has begun some~~where in~~ its logical middle. It usually begins with observations. Then it is noticed that some of these observed statements are logical consequences of others. After attempts to organize these statements deductively, someone finally conjectures hypotheses which would imply conclusions corresponding to these observations. It may happen that these hypotheses imply other conclusions that may be checked against observation. If some of these do not check then we must abandon or modify our hypotheses; that is we try to imagine a

new set of postulates whose implications will check with experience when interpreted in some concrete way. Thus empirical science poses problems to pure mathematical science. On the other hand, pure mathematical science deduces theorems logically which have to be tested experimentally by empirical science. That is, deduction may suggest crucial experiments to be performed in the laboratory; and laboratory observations may suggest programs of deductive thought. The two play mutually helpful and complementary roles.

Complete * logical organization, such as has been achieved in geometry, for example, is the final stage in the development of a science. If you say that it is extremely difficult to apply such standards of logical rigor to, say, the social sciences, because they are more complex than elementary mathematics, we must agree that it is difficult. It was also difficult to introduce successful logical theories into the apparent chaos of astronomy but it has been done with remarkable success in the course of many years of concentrated and painstaking effort. Today, chemistry and biology are beginning to emerge from their early stages of development and are taking shape as genuine logical systems. Small beginnings are being made in this direction in such subjects as psychology and economics, which are, compared to physics and mathematics, in their early infancy.

Briefly, the goal of science is to find assumptions, the fewer and simpler the better, whose logical consequences correspond with experience. Thus the first attempts to "explain" physical phenomena were mythologies. If each observed phenomenon is "explained" by saying that it is due to the whim of spirits, there is no way to disprove this theory. But it is a useless theory since it does not enable you to predict future phenomena. It must not be supposed that anything said here detracts from the great achievements of science. On the contrary, it should help you to appreciate the remarkable success of scientists in finding postulates whose logical consequences correspond with the varied observations of experience in so many diverse directions; that is, in finding abstract mathematical sciences whose concrete interpretations correspond closely with observation. If little has been

* According to present standards. This should be understood whenever the expression "completely logical" is used.

said here about the details of these scientific achievements, it is partly because of lack of space, and partly because it is assumed that you are at least vaguely familiar with them. Therefore we have concentrated on explaining the postulational nature of science, which you may never have understood.

In any case, an understanding of the nature of abstract postulational thinking and its concrete applications, and of truth and validity, should help us in being critical of our beliefs, and in finding out the assumptions and reasoning on which our beliefs rest. It should make us less dogmatic and more tolerant, since it destroys the assurance, that some of us have, that we know with certainty the true answers and the only true answers to all questions.

10. Generality and abstractness. In arithmetic one begins by counting apples or oranges. The importance of the abstract "number" comes into play when one realizes that three apples, three oranges, or three objects of any nature whatever have some abstract property in common, the number three. Thus instead of making the particular statment that two apples plus one apple make three apples, one prefers the general abstract statement $2 + 1 = 3$.

Later on, one observes that $9 - 4 = 5 \cdot 1$ or $3^2 - 2^2 = (3 + 2)(3 - 2)$, and $16 - 4 = 6 \cdot 2$ or $4^2 - 2^2 = (4 + 2)(4 - 2)$, and $25 - 4 = 7 \cdot 3$ or $5^2 - 2^2 = (5 + 2)(5 - 2)$. The mathematician naturally prefers the general statement $x^2 - 4 = (x + 2)(x - 2)$ which holds for *all* numbers x, instead of the particular statements about individual numbers made above. In fact, one prefers still more the general statement $x^2 - y^2 = (x + y)(x - y)$ which holds for *all* numbers x and *all* numbers y.

Similarly, in geometry one prefers the general statement that "the area of every rectangle is equal to its base multiplied by its altitude," instead of any particular statement such as "the area of a rectangle whose base is 2 inches and altitude 3 inches is 6 square inches."

In fact, this is characteristic of science in general. The scientist does not want to have to invent a separate explanation for each individual phenomenon. He much prefers to find a general law that will explain all of a large class of phenomena.

No two phenomena can ever be exactly alike since they differ at least in time or place. Hence generality can be obtained only by abstracting from the phenomena those characteristics which they have in common, ignoring other characteristics. Similarly, in the beginnings of arithmetic one learns to ignore the color, taste, etc., of the three apples or three oranges and to consider only the common characteristic of number. The same situation arises in logic when one ignores the content of the statements in an argument and considers only their form (see section 5). Needless to say, generalizations are often discovered or suggested by observation or inductive reasoning.

Note carefully that in mathematics the word "generality" does not connote vagueness, as it sometimes does in everyday usage. Nor does "abstractness" imply having one's feet firmly planted on a cloud. A general result, in mathematics, always includes within it all the special results that may have suggested it. And an abstract statement yields, upon appropriate application to specific subject matter, all the concrete statements of which it is the prototype. It may help the student to understand the development and importance of mathematics if he grasps the significance of generality and abstractness. The desire to arrive at general statements about all of some class of things often actuates the mathematician and gives direction to his research.

EXERCISES

1. Explain the nature of (a) inductive logic; (b) deductive logic.

2. Is the conclusion of a valid argument always true? Explain the distinction between truth and validity.

3. Explain the nature of postulational thinking. What is an abstract mathematical science? A concrete interpretation? Define and discuss the relationship between pure and applied mathematics.

4. Explain the bearing of your answers to parts (c) and (d) of exercise 14, section 4, on the acceptance or rejection of scientific theories or hypotheses.

5. Make up examples of different hypotheses which imply the same conclusion.

REFERENCES

E. T. Bell, *The Queen of the Sciences*, G. E. Stechert, N. Y., 1938.
E. T. Bell, *The Development of Mathematics*, McGraw-Hill, N. Y., 1940.
A. A. Bennett & C. A. Baylis, *Formal Logic*, Prentice Hall, N. Y., 1939.

M. R. Cohen & E. Nagel, *An Introduction to Logic and Scientific Method*, Harcourt Brace, N. Y., 1934.

A. Einstein & L. Infeld, *The Evolution of Physics*, Simon & Schuster, N. Y., 1938.

C. J. Keyser, *Thinking About Thinking*, E. P. Dutton, N. Y., 1926.

C. J. Keyser, *Pastures of Wonder*, Columbia University Press, N. Y., 1929.

C. J. Keyser, *The Human Worth of Rigorous Thinking*, Columbia University Press, N. Y., 1925.

B. Russell, *Mysticism and Logic* (Chapters II and IV), Longmans Green, London, 1929.

A. N. Whitehead, *An Introduction to Mathematics* (Chapter I), Henry Holt, N. Y., 1911.

A. N. Whitehead, *Science and the Modern World* (Chapter II), Macmillan, N. Y., 1925.

Chapter III

THE SIMPLEST NUMBERS * †

11. Introduction. We did not attempt in section 8 to give a complete set of postulates for Euclidean geometry and to deduce the familiar theorems from them. Such a task involves considerable technical difficulty. Neither your high school text in geometry, nor Euclid's original work, *The Elements* (about 300 B.C.), succeeded in presenting a completely logical deduction of the theorems from the postulates according to present standards. In fact, such a treatment could not be given without a more complete set of postulates, and was not given until the last years of the 19th century. However, the treatment of geometry in your high school course, though incomplete, did make clear that the subject may be considered as an example of postulational thinking in the light of our discussion in section 8. We may now turn to algebra and arithmetic, which we shall treat in greater detail because your high school treatment of these subjects was much inferior, from our point of view, to that of your geometric studies. Much of your algebra was probably learned by rote; here, it will be treated logically. While we cannot attempt a complete treatment because of technical difficulties, which would tax the abilities of more advanced students than you, we shall try to do enough to convince you that algebra has a logical structure similar to that of geometry, and to acquaint you with its historical growth. Since a completely logical treatment is not feasible here, we shall have to ask you to adopt a kind of split personality.

* *Note to the student:* This chapter should be read slowly since the idea of logical proof in connection with numbers is strange to you. After completing this chapter, the work of Chapters IV, and, especially, V will be found to be much easier.

† *Note to the instructor:* If it is felt that fewer proofs than are given here will suffice to enable the student to grasp the idea of logical proof, then the remaining proofs may be omitted and the facts assumed or discussed informally without disturbing the continuity of the book. We have tried to include as much rigor as anyone would want since it is more convenient for the instructor to omit than to insert such details.

Sometimes we shall address you as if you were children or barbarians and we shall build up with careful logic even the most elementary of subjects. At other times we shall ask you to recall some simple things with which you have long been familiar, such as the addition and multiplication tables.

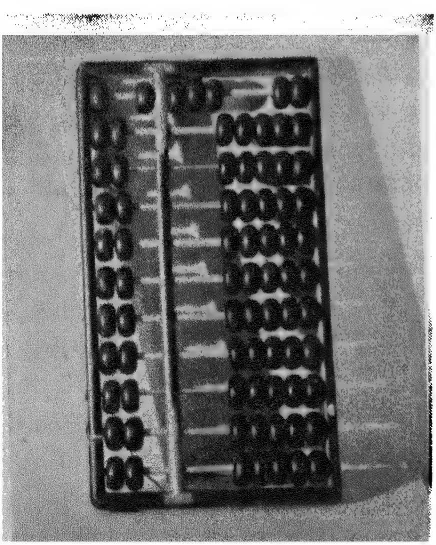
The system of numbers, with which you became acquainted in your previous study of algebra, did not come into being full-grown. It developed very slowly, beginning in prehistoric times and reaching its present state only recently. The so-called natural numbers, 1, 2, 3, 4, and so on, certainly prehistoric in origin, were doubtless the only numbers known to the human race for a long time. Fractions, like 3/5, 7/2, 1012/357, etc. were probably invented in order to facilitate the handling of problems of measurement concerning the division of things (like real estate or harvests) into equal parts. In their study of geometric magnitudes the Greeks discovered that some lengths could not be represented even by fractions (see section 24). Considerations of this sort led to the invention of irrational numbers like $\sqrt{2}$. When it became convenient, for various purposes, to be able to subtract a larger number from a smaller one, negative numbers were invented. If one wishes to take the square root of a negative number, one is led to invent the so-called imaginary numbers like $\sqrt{-2}$. Finally, there was evolved the system of so-called complex numbers, like $2 + 3\sqrt{-7}$, with which you became acquainted in school. In the present chapter we shall discuss only the simplest numbers, namely the natural numbers and fractions, in some detail. In Chapter IV we shall continue our discussion of the gradual evolution of the number system. Since our purpose is not merely historical but also to develop the subject from a modern logical point of view, we shall depart somewhat from the chronological order of development, for pedagogical reasons. It would be foolish to retrace here all the blunderings of the human race in its efforts to create a satisfactory number system.

In Chapter V, we shall indicate how the familiar manipulations of algebra might be based on a set of postulates and undefined terms. However, as remarked in Chapter II, our choice of postulates is usually suggested by some sort of experience. It is the purpose of Chapters III and IV to acquaint you with experiences of the human race which lead up to the formulation of the postulates used

in Chapter V, and, at the same time, to introduce you gradually to the use of logical proofs and strict definitions in connection with numbers. Chapters III and IV do not themselves constitute an abstract mathematical science, since the basic term *natural number* will not be regarded here as an undefined term devoid of meaning, but will rather have a concrete meaning, understood intuitively, to be indicated at the beginning of the next section.

Many of the things we shall write down may seem trivial to you but you must recall that, as far as possible, you are to remember nothing of your high school algebra, that we want to write down everything that we take for granted, and that we must use nothing in our proofs except what is explicitly assumed or previously proved.

12. Natural numbers. Addition and equality. Let us begin by imagining ourselves to be young children or primitive savages who know nothing of numbers. Our earliest mathematical discovery would probably be that two apples, two fingers, two people, two dogs, etc., have something in common. Thus, the first numbers to be invented would probably be one and two. Some primitive races are said to find it too much of a strain to enter into further refinements and have in their language only three words of number—one, two, and many. Since none of us can grasp intuitively large numbers of objects like the number of blades of grass on a lawn or the number of people in an auditorium, it is likely that our ancestors would not have progressed much further than this were it not for the invention of counting. However, having invented the process of counting objects in succession (that is, "counting by ones"), our ancestors, early in prehistoric times, surely developed the numbers 1, 2, 3, 4, and so on,



Chinese abacus or swan-pan
FIG. 13

which are used in counting. These are called the **natural numbers.** From our experience with counting objects we develop next the idea of the sum of two numbers. For example, if we have three apples and we add two more apples we have five apples; this fact leads to the statement $3 + 2 = 5$. We then find it convenient to record and memorize the resulting tables of addition

$$\begin{array}{llll} 1+1=2 & 2+1=3 & 3+1=4 & \cdots \\ 1+2=3 & 2+2=4 & 3+2=5 & \cdots \\ 1+3=4 & 2+3=5 & 3+3=6 & \cdots \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \end{array}$$

which we can use in calculating sums. Devices other than tables, such as the abacus (Fig. 13) or various forms of finger-reckoning, were once used universally for addition and are still used today among unlettered people. We shall now assume that you are familiar with the usual way of writing numbers, and that you can add. We also assume that you are familiar with the use of letters to represent unspecified numbers.

The intuitive fact that any two natural numbers can be added to obtain another natural number may be formulated as a postulate or assumption as follows.

A_1. *Given any pair of natural numbers, a and b, in the stated order, there exists one and only one natural number, denoted by $a + b$, called the* **sum** *of a and b.* (This is called the **postulate of closure for addition of natural numbers.**)

The significance of the word closure is that the natural numbers form a "closed" system under addition in the sense that the result of adding two numbers in the system always remains in the system. It is as though all natural numbers filled a fenced-in enclosure and performing the operation of addition never takes us outside the enclosure. This may be made clearer by considering the following examples of systems not closed under addition.

Example 1. Consider the system of all natural numbers from 1 up to 10, inclusive. This system is not closed under addition for it would not be correct to say that given any two numbers in

the system there is a number in the system called their sum. For instance, 4 and 7 are in the system but their sum is not.

Example 2. Consider the system of all odd numbers; that is, $1, 3, 5, 7, \cdots$.* This system is not closed under addition since the sum of two odd numbers is even. On the other hand the system of all even numbers $(2, 4, 6, \cdots)$ is closed under addition for the sum of two even numbers is always even. (These statements are not being proved here; we are merely taking these illustrations intuitively.)

If we add two pebbles to a collection of three pebbles we get the same number of pebbles as if we had added three pebbles to a collection of two pebbles. Thus $3 + 2 = 2 + 3$. This, and similar experience, suggests formulating the following property of natural numbers as a postulate.

A_2. *If a and b are any natural numbers, $a + b = b + a$.* (Called the **commutative law for addition of natural numbers.**)

That is, we may commute or interchange the order of the terms in the sum $a + b$.

Let us recall that parentheses are used to group together what is in them. This is merely a rule or agreement in the grammar of our written language of algebra. Thus $2 + (3 + 4)$ means that we add 3 and 4 first, obtaining 7, and then find $2 + 7$, obtaining 9. Similarly, $(2 + 3) + 4$ means $5 + 4$ or 9. The fact that we get the same result from both expressions, although the operations performed are different, leads us to formulate the following property of natural numbers as a postulate.

A_3. *If a, b, c are any natural numbers, $(a + b) + c = a + (b + c)$.* (Called the **associative law for addition of natural numbers.**)

This says that we get the same result whether we add the sum of a and b to c, or a to the sum of b and c. That is, the parentheses may be put around the first or the second pair of natural numbers, at will.

To say that $a = b$ means that a and b are different symbols representing the same natural number; that is, they are different names for the same thing. If $a = b$, then, in any statement involving a, we may replace a by b and vice versa. This is a fundamental princi-

* Wherever three dots occur, the first dot means "and," the second dot means "so," and the third dot means "on."

ple of our underlying logic, called the **principle of substitution.** For example, postulate A_3 tells us that when we see the expression $(a + b) + c$ we may replace it by the expression $a + (b + c)$. Similarly A_2 tells us that we may substitute $b + a$ for $a + b$. The principle of substitution is related to the following postulates concerning equality, which we shall use, often without explicit mention.

E_1. *If a is a natural number, $a = a$. (Law of identity.)*

E_2. *If a and b are any natural numbers and $a = b$, then $b = a$.*

E_3. *If a, b, c are natural numbers, and if $a = b$ and $b = c$, then $a = c$. (Things equal to the same thing are equal to each other.)*

E_4. *If $a = b$ and $c = d$, all letters representing natural numbers, then $a + c = b + d$. (If equals are added to equals, the results are equal.)*

As a consequence of E_3 (or substitution) we may use long chains of equalities; thus if $a = b$, $b = c$, $c = d$, $d = e$, $e = f$, we may conclude that $a = f$, and we may write $a = b = c = d = e = f$.

A statement of equality is called an **equation;** the expressions on either side of the equals sign are referred to as the **left member** and **right member** of the equation, respectively.

We can now prove some simple theorems like the following.

THEOREM 1. *If x, y, and z are any natural numbers, then $(x + y) + z = (z + x) + y$.*

Proof. By hypothesis x and y are natural numbers. Hence, by the postulate of closure for addition of natural numbers (A_1) it follows that $x + y$ is a natural number. That is, $x + y$ is a long name for a certain natural number. By hypothesis z is a natural number. By the commutative law for addition of natural numbers (A_2), applied to the natural numbers $(x + y)$ and z, we have

(1) $$(x + y) + z = z + (x + y).$$

By the associative law for addition of natural numbers (A_3), applied to the right member of (1), we have

(2) $$z + (x + y) = (z + x) + y.$$

From (1) and (2) we obtain by substitution, or by E_3, the result

$$(x + y) + z = (z + x) + y$$

which is what we had to prove.

Note that before we applied A_2 to the symbols $(x + y)$ and z, we had to be sure that they both represented natural numbers, since A_2 applies only to natural numbers. That $(x + y)$ was a natural number was established by using A_1.

Remark 1. The sum of three or more terms has not been determined by A_1. Nor does the addition table tell you the sum of more than two terms. Thus a symbol like $a + b + c$ has no meaning at present. But $(a + b) + c$ has a meaning and so has $a + (b + c)$ since each indicated step in either expression is addition of two terms which is possible by A_1. Furthermore by A_3, both of these expressions are equal. Hence we *define* $a + b + c$ to mean the number obtained by inserting parentheses in either way: $a + (b + c)$ or $(a + b) + c$. As for sums involving four terms, we observe that $([3 + 4] + 6) + 5 = (7 + 6) + 5 = 13 + 5 = 18$ while $[3 + 4] + (6 + 5) = 7 + 11 = 18$. This suggests that we can now prove theorems like the following, on the basis of our postulates.

THEOREM 2. *If a, b, c, and d are any natural numbers then* $([a + b] + c) + d = [a + b] + (c + d)$.

Proof. By hypothesis, a and b are natural numbers. Hence the postulate of closure for addition of natural numbers (A_1) tells us that $[a + b]$ is a natural number. Since $[a + b]$, c, and d are now known to be natural numbers, the associative law for addition of natural numbers (A_3) tells us that $([a + b] + c) + d = [a + b] + (c + d)$. This is what we had to prove.

Similarly we could show that all expressions obtained by inserting parentheses in the various possible ways must yield the same result. Thus $[a + (b + c)] + d = a + [(b + c) + d]$, and so on. Hence we define the expression $a + b + c + d$ to mean the number arrived at by pairing off the terms by inserting parentheses in any way that makes sense, since all such ways yield the same result. In general, the sum of any number of terms $a + b + c + \cdots + k$ may be defined, and we can write such sums without parentheses, or we may group the terms in any way by putting in parentheses at will (as long as we are only adding). This could be proved rigorously * but we will not go into details here. *We*

* With the help of further postulates not stated here. In particular, the postulate of mathematical induction, discussed in Chapter XIV, is needed.

shall say that such insertion or removal of parentheses in sums is justified by the ***generalized associative law for addition of natural numbers.***

Remark 2. The commutative and associative laws may seem trivial, but there are in fact many operations which do not obey them. To anticipate, subtraction obeys neither. For $2 - 7$ is not the same as $7 - 2$; and $(8 - 4) - 2$ is not the same as $8 - (4 - 2)$, since the first expression means $4 - 2$ or 2 and the second means $8 - 2$ or 6. Similarly the operation of doing one thing after another is not commutative, for the order in which you do two things may well matter; for example, shelling an egg and beating it up; or undressing and bathing. Chemistry students all know that one may safely pour sulphuric acid into water, while it is dangerous to pour water into sulphuric acid.

Remark 3. The student should memorize the names of these laws, not their numbers, although for the sake of brevity we shall often use the numbers for cross-references. When quoting a postulate as justification for a step in a proof, the student should either write out the postulate in full, or use abbreviations like Comm. Add. Nat. Nos. for the commutative law for addition of natural numbers.

Remark 4. The commutative and associative laws are implicitly used when we say that we can add a column of figures either up or down. Thus adding the column

$$\begin{array}{r} 7 \\ 4 \\ \underline{6} \\ 17 \end{array}$$

up means $(6 + 4) + 7$ while adding it down means $(7 + 4) + 6$. The results are the same.

EXERCISES

1. Name the postulate which justifies each of the following statements, all letters representing natural numbers:

(*a*) $5 + 7 = 7 + 5$.
(*b*) $(5 + 7) + 2 = 5 + (7 + 2)$.
(*c*) $x + y = y + x$.
(*d*) $(m + x) + r = m + (x + r)$.
(*e*) $a + b$ is a natural number.

2. Prove the following, justifying each step by means of one of our postulates, all letters representing natural numbers. Do not use the generalized associative law discussed in remark 1.

(*a*) $[a+(b+c)]+d$ is a natural number. (Hint: use A_1 several times.)
(*b*) $a+[(b+c)+d]$ is a natural number.
(*c*) $[a+(b+c)]+d = a+[(b+c)+d]$.
(*d*) $(x+y)+z = x+(z+y)$. (Hint: use A_3 and A_2.)
(*e*) $(x+y)+z = y+(x+z)$.
(*f*) $(a+b)+c = c+(a+b)$. (Hint: use A_1 and A_2.)
(*g*) $(a+b)+(c+d) = (c+d)+(a+b)$.
(*h*) $[a+(b+c)]+d = a+[d+(b+c)]$. (Hint: use A_1, A_3, and A_2.)
(*i*) $[a+(b+c)]+d = (a+b)+(c+d)$.
(*j*) $(x+y)+z = (x+z)+y$.

13. Multiplication of natural numbers. Early in the history of the human race it must have been noticed that repeated additions of the same natural number can be abbreviated by the invention of multiplication. Thus, a herdsman counting cattle passing through a gate three abreast might have observed that he could "count by threes" more quickly than by "ones," and he might have then decided to remember that four threes add up to twelve, and similar facts. Hence we might define "4 times 3" to mean $3 + 3 + 3 + 3$. This leads us to the following definition.

DEFINITION 1. *If a and b are natural numbers, the* **product** *of a and b shall mean the number* $b + b + b + \cdots + b$ *where there are a terms in the sum. In symbols,* $ab = b + b + \cdots + b$ *(a terms on the right of the equals sign). The numbers a and b are called the* **factors** *of the product. The product of a and b will be written* * *as ab or* $a \cdot b$. *In particular* $1 \cdot b = b$.

For example, $4 \cdot 3 = 3 + 3 + 3 + 3$. Note that our definition of product involves only terms (like "sum") which have already been introduced. Since the sum of any number of natural numbers is a natural number, it is natural to assume the following postulate.

M_1. *If a and b are any two natural numbers, given in the stated order, there exists one and only one natural number denoted by ab or* $a \cdot b$, *called the product of a and b.* (This is called the **postulate of closure for multiplication of natural numbers.**)

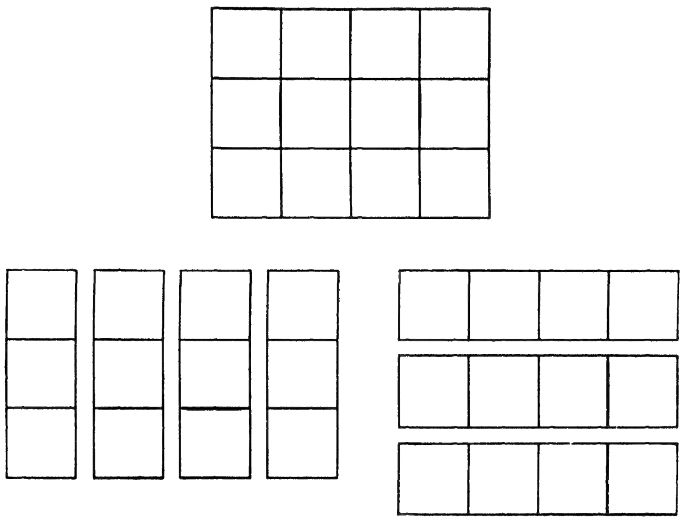
This could have been proved † but we shall assume it here. An

* The dot is always used when specific numbers are multiplied. For example 4 times 3 is written $4 \cdot 3$, not 43.

† See footnote, page 48.

example of a system of numbers not closed under multiplication is the system of all natural numbers between one and ten inclusive; for the numbers 3 and 4 are in this system but their product is not. Postulate M_1 asserts that the system of *all* natural numbers *is* closed under multiplication.

According to definition 1, $4 \cdot 3 = 3 + 3 + 3 + 3$ while $3 \cdot 4 = 4 + 4 + 4$. These expressions represent different operations but



$4 \cdot 3 = 3 \cdot 4$

FIG. 14

the result is the same in both cases. Our experience suggests that we should always get the same result even if the order of our factors is reversed. Therefore we assume the following postulate.

M_2. *If a and b are any natural numbers, then $ab = ba$.* (This is called the **commutative law for multiplication of natural numbers.**)

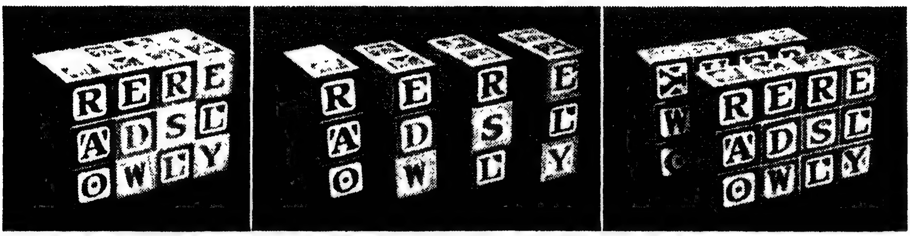
Similarly our experience tells us that although $(4 \cdot 3) \cdot 2 = 12 \cdot 2$ and $4 \cdot (3 \cdot 2) = 4 \cdot 6$ represent different operations, they yield the same result. Therefore it is natural to make the following postulate.

M_3. *If a, b, c, are any natural numbers, then $(ab)c = a(bc)$.* (This is called the **associative law for the multiplication of natural numbers.**)

From these postulates we can deduce some simple theorems like the following.

THEOREM. *If x, y, z are any natural numbers, then $x(yz) = z(xy)$.*

Proof. By M_3, $x(yz) = (xy)z$. By M_1, (xy) is a natural number. Hence we may apply M_2 to the natural numbers (xy) and z, obtaining $(xy)z = z(xy)$. By E_3 or substitution we have $x(yz) = z(xy)$ which was to be proved.



$4 \cdot (3 \cdot 2) = 2 \cdot (4 \cdot 3)$

FIG. 15

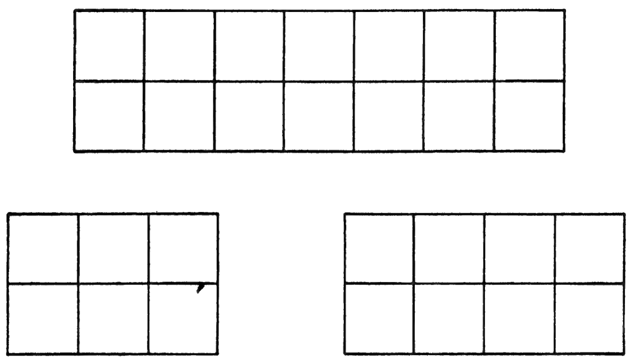
We now assume you to be familiar with the tables of multiplication which can be derived from definition 1 and the tables of addition:

$1 \cdot 1 = 1$	$2 \cdot 1 = 2$	$3 \cdot 1 = 3$	$\cdots$
$1 \cdot 2 = 2$	$2 \cdot 2 = 4$	$3 \cdot 2 = 6$	$\cdots$
$1 \cdot 3 = 3$	$2 \cdot 3 = 6$	$3 \cdot 3 = 9$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	

Remark 4. The product of three terms is not defined yet. But M_3 enables us to define and use symbols like abc or $abcd$. *We shall insert or remove parentheses in a product of any number of factors freely, saying that this is justified by the* ***generalized associative law for multiplication.*** Compare what we said in connection with addition in Remark 1, section 12.

We recall the convention of notation (a rule of grammar in the written language of algebra) that *in a chain of additions and multiplications the multiplications are to be done first except where otherwise indicated by parentheses.* This convention enables us to omit cumbersome parentheses around products. For example $2 + (3 \cdot 4)$ can be written simply as $2 + 3 \cdot 4$. Note that $2 + 3 \cdot 4$ means $2 + 12$ or 14 while $(2 + 3) \cdot 4$ means $5 \cdot 4$ or 20. Simi-

larly, $2 \cdot 3 + 4 = 6 + 4 = 10$ while $2 \cdot (3 + 4) = 2 \cdot 7 = 14$. However, we observe that $2 \cdot (3 + 4) = 2 \cdot 3 + 2 \cdot 4$ since the left member is $2 \cdot 7$ while the right is $6 + 8$. In fact, $2 \cdot (3 + 4) = (3 + 4) + (3 + 4)$ by definition of product. But, using the asso-



$2(3 + 4) = 2 \cdot 3 + 2 \cdot 4$

FIG. 16

ciative and commutative laws we have $(3 + 4) + (3 + 4) = 3 + 4 + 3 + 4 = 3 + 3 + 4 + 4 = (3 + 3) + (4 + 4) = 2 \cdot 3 + 2 \cdot 4$. This suggests the following general postulate.

D. If a, b, c are natural numbers, then $a(b + c) = ab + ac$. (This is called the **distributive law for natural numbers.**)

The effect of the multiplier a is distributed between the terms b and c. This is the first postulate we have made which connects addition and multiplication. In high school this law was referred to as "removing parentheses" if read from left to right, and as "taking out a common factor" if read from right to left.

We also assume the following postulate of equality.

E_5. If $a = b$ and $c = d$, all letters representing natural numbers, then $ac = bd$. (If equals are multiplied by equals, the results are equal.)

EXERCISES

1. Name the postulate or postulates which justify each of the following statements, all letters representing natural numbers:

(*a*) $3 \cdot 5 = 5 \cdot 3$. (*b*) $(3 \cdot 5) \cdot 4 = 3 \cdot (5 \cdot 4)$.
(*c*) $3(5 + 4) = 3 \cdot 5 + 3 \cdot 4$. (*d*) $5 \cdot 2 + 5 \cdot 3 = 5(2 + 3)$.
(*e*) $xy = yx$. (*f*) $(xy)z = x(yz)$.
(*g*) $x(y + z) = xy + xz$. (*h*) $mp + mq = m(p + q)$.

2. According to the definition of multiplication what is the meaning of $3 \cdot 5$? of $5 \cdot 3$? of $5a$?

3. Find the value of each of the following:

(*a*) $3 + 4 \cdot 5$. (*b*) $(3 + 4) \cdot 5$. (*c*) $3 \cdot 4 + 5$.
(*d*) $3 \cdot (4 + 5)$. (*e*) $3 + 4 \cdot 5 + 6$. (*f*) $(3 + 4) \cdot 5 + 6$.
(*g*) $(3 + 4) \cdot (5 + 6)$. (*h*) $3 + 4 \cdot (5 + 6)$.

4. Notice that $2 + 3 \cdot 4 = 2 + 12 = 14$, and $2 \cdot (3 + 4) = 2 \cdot 7 = 14$. Hence $2 + 3 \cdot 4 = 2 \cdot (3 + 4)$. May we conclude from this that $a + b \cdot c = a \cdot (b + c)$ for all natural numbers a, b, and c?

5. Prove each of the following, justifying each step by means of one of our postulates, all letters representing natural numbers.

(*a*) $(xy)z = x(zy)$. (Hint: use M_3 and M_2.)
(*b*) $(xy)z = y(xz)$.
(*c*) $(xy)(zu) = (zu)(xy)$. (Hint: use M_1 and M_2.)
(*d*) $(xy)z = (xz)y$.
(*e*) $xy + xz$ is a natural number. (Hint: use M_1 and A_1.)

14. Some theorems about natural numbers. We are now in a position to prove some theorems about natural numbers on the basis of our postulates. The student's chief difficulty here will be the necessity for confining himself to steps which are justified by the postulates. It will be necessary to refrain consciously from performing manipulations, made familiar by habitual use, which are not assumed and have not yet been proved to be permissible here. We give two illustrations.

THEOREM 1. *If x, y, z are any natural numbers, then $(y + z)x = yx + zx$.*

Note that this is not the distributive law as we assumed it, since the number outside the parenthesis here is on the right instead of the left of the parenthesis. It will be advisable to restate all theorems indicating hypothesis and conclusion clearly, as follows.

Hypothesis. x, y, and z are natural numbers.

Conclusion. $(y + z)x = yx + zx$.

Proof. By the postulate of closure for addition of natural numbers (A_1), $y + z$ is a natural number. Now x is a natural number by hypothesis. Hence, applying the commutative law for multiplication of natural numbers (M_2) to the natural numbers $(y + z)$ and x, we have

$$(y + z)x = x(y + z).$$

But, by the distributive law (D),

$$x(y + z) = xy + xz.$$

Now, by the commutative law for multiplication (M_2),

$$xy = yx \text{ and } xz = zx.$$

Hence, $\quad xy + xz = yx + zx \quad$ (by E_4).

Finally, $\quad (y + z)x = yx + zx \quad$ (by E_3 or substitution).

This is what we had to prove.

Remark. Instead of recognizing that $y + z$ may be regarded as a long name for a natural number, the student may substitute a short name for it if he wishes. Thus we might have said: let $y + z = a$. Then $(y + z)x = ax$ by substitution. Now $ax = xa$ by M_2. Hence $(y + z)x = x(y + z)$ by substitution. The rest of the proof would proceed as before. However, this substitution is unnecessary. Note that the step "let $y + z = a$" is not a step of reasoning but merely a matter of name-calling. Logic should not be confused with name-calling even in a political campaign.

THEOREM 2. *If a, b, c, d are natural numbers, then $a(b + c + d) = ab + ac + ad$.*

Hypothesis. a, b, c, d are natural numbers.

Conclusion. $a(b + c + d) = ab + ac + ad$.

Proof. By the associative law for addition of natural numbers (see Remark 1, section 12), $b + c + d$ can be written as $[b + c] + d$. By A_1, $[b + c]$ is a natural number. Then

$$a(b + c + d) = a([b + c] + d) \quad \text{(by substitution).}$$

Now, $\quad a([b + c] + d) = a[b + c] + ad \quad$ (by D).

But, $\quad a[b + c] = ab + ac \quad$ (by D).

Thus, $\quad a[b + c] + ad = ab + ac + ad \quad$ (by substitution).

Finally, $\quad a(b + c + d) = ab + ac + ad \quad$ (by E_3 or substitution.)

Notice that the steps in these proofs are things that you would have done in high school without stopping to think at all. Our purpose is not to "get the answer" quickly, but rather to deduce our results from the postulates or assumptions we set down, being

careful to justify each step on that basis. The extreme care we have employed in justifying each minute step logically is doubtless new to you. All of mathematics from these small beginnings to the most advanced branches can be built up in this way with each step justified by a logically rigorous argument.

While you will have had little difficulty in following the proofs given above, you may be at a loss as to how to proceed in the original exercises below. When the proof is given, you can see that each step is logically justified. But when you have to invent your own proof you may be puzzled as to which of the many possible justifiable steps you ought to take. The best counsel that the author can think of in this connection is the advice given to Alice by the Cheshire Cat in *Alice in Wonderland* * by C. L. Dodgson (Lewis Carroll). Alice asks the Cat:

' "Would you tell me, please, which way ought I go from here?"

' "That depends a good deal on where you want to get to," said the Cat.

' "I don't much care where—," said Alice.

' "Then it doesn't matter which way you go," said the Cat.'

Similarly, it is usually easier to see what to do next if you know what goal you are driving at. Thus when you are trying to determine which of many possible logically justifiable steps you should take, think of what you are trying to prove and select some justifiable step which will bring you nearer to that goal.

As an illustration, let us analyze theorem 1 so as to discover naturally how the proof should go. We have to prove that $(y + z)x = yx + zx$. There is only one postulate which deals with both addition and multiplication together, namely the distributive law. But we cannot apply the distributive law since x is on the wrong side of the parenthesis. Therefore we would like to put the x on the left side of the parenthesis. This could be done by the commutative law for multiplication provided the parenthesis represented a natural number. But it does by the law of closure for addition. Therefore we begin our proof by saying that $y + z$ is a natural number by A_1. Then by M_2, we get $(y + z)x = x(y + z)$. By D, $x(y + z) = xy + xz$. Now we have

* This book, and *Through the Looking Glass*, by the same author, make good reading for adults. The Reverend Charles L. Dodgson was a teacher of mathematics, and his tales contain many sly allusions of mathematical and philosophical character.

attained our first objective of removing the parentheses. But the right member of the last step is not yet what we have to prove. But it is easily made so by the commutative law for multiplication. Having thus analyzed the proof, we then write it down in good order as above.

Theorems 1 and 2 and exercises 5 and 6 below, are all instances of what may be called the **generalized distributive law.** That is,

$$a(b + c + d + \cdots + k) = ab + ac + ad + \cdots + ak$$

and $(b + c + d + \cdots + k)a = ba + ca + da + \cdots + ka$

no matter how many terms are within the parenthesis. We shall not prove this generalized distributive law here but we shall use it in later sections. Do not use it in the following exercises.

Remark 1. In applying any postulate or theorem to any expression, or in using "substitution," always be sure you know what you are substituting for what. Write out all "reasons" in full. If you are doubtful about whether a certain postulate applies to the situation in hand, it is a good idea to write out the postulate and then substitute in it the expressions which are to play the part of the symbols in the postulate. For example, in the above analysis of theorem 1 we want at one stage to use M_2 which says: if a and b are natural numbers, then $ab = ba$. Applying this to the situation in hand, we write: if $(y + z)$ and x are natural numbers then $(y + z)x = x(y + z)$. This will be all right provided $(y + z)$ *is* a natural number. And so on.

Remark 2. The student should note that every step in our proofs, no matter how small, is carefully justified. However, if this practice were continued throughout the book, the proofs would become intolerably long. Therefore, as a matter of expedience, as soon as we have become sufficiently familiar with the justification of tiny steps such as those involving the commutative and associative laws, we shall begin to slur over them; that is, we shall begin to do several of them simultaneously and finally shall do them mentally without explicit mention. However, the student should not relax his standard of proof in this way until he has become thoroughly familiar with the task of proving things in full, justifying even the smallest steps. He should be *capable* of putting in *every* step if it is desired.

EXERCISES

(*a*) *Prove the following, all letters being understood to represent natural numbers, justifying each step by means of a postulate or a previously proved theorem.* (*b*) *Also illustrate each formula by replacing the letters by particular numbers.*

1. $m(p + q) = qm + mp.$

2. $x(y + z) = zx + yx.$

3. $(a + b)c = ca + cb.$

4. $(m + n)t = tn + mt.$

5. $a(b + c + d + e) = ab + ac + ad + ae.$ (Hint: write $b + c + d + e = [b + c] + [d + e]$ and proceed as in Theorem 2.)

6. $a(b + c + d + e + f) = ab + ac + ad + ae + af.$ (Hint: write $b + c + d + e + f = [b + c] + [d + e] + f$ and apply Theorem 2.)

7. $(a + b)(c + d) = ac + bc + ad + bd.$ Compare the way you would have obtained this result in high school.

8. $(a + b)(c + d + e) = ac + bc + ad + bd + ae + be.$

9. $2(x + 3) = 2\,x + 6.$

10. $(x + 5) \cdot 3 = 3\,x + 15.$

11. $(a + 2)(b + 3) = ab + 3\,a + 2\,b + 6.$

12. $(x + 5)(y + 2) = 5\,y + xy + 2\,x + 10.$

15. Subtraction and division of natural numbers. "My friend has more marbles than I have; how many more?" This question occurs early in a child's life just as it must have occurred early in the history of the human race. It leads to the concept of subtraction. Children are often taught to subtract 2 from 5 by asking, "What must be added to 2 in order to get 5?" Formalizing this question, we make the following definition.

DEFINITION 1 (*a*). *If a and b are natural numbers then the symbol $a - b$ means a natural number x such that $b + x = a$, provided such a number x exists. The symbol $a - b$ is called the* ***difference*** *a minus b.*

This is not unnatural since the number of marbles which must be added to 2 in order to get 5 is clearly the same as the number of marbles left when 2 are taken away from 5. This is exhibited in the fact that there are two schools of thought among shopkeepers in regard to making change from a five-dollar bill. Some will figure out what is left when the purchase price is taken away from \$5; others will start with the purchase price and calculate what must be added to it to obtain \$5.

Definition 1 (*b*). *If there exists a natural number x such that $b + x = a$, then a is said to be* ***greater than*** *b or b is said to be* ***less than*** *a; in symbols, $a > b$ or $b < a$.*

Note that, as always, we define new terms in terms of old ones. Thus "$a - b$" and "$a > b$" are defined in terms of addition.

Notice that while sums and products of two natural numbers a and b always exist (among the natural numbers), the difference may not. For example, $5 - 2$ means a number x such that $2 + x = 5$; or $x = 3$ since we discover from the addition table that $2 + 3 = 5$. But $2 - 5$ means a number x such that $5 + x = 2$, and no such number exists among the natural numbers. Hence $2 - 5$ is at present a meaningless symbol. Our definition 1(a) is like the requirements for a degree or title. If a number x satisfies the requirement $b + x = a$, then x is entitled to be called $a - b$. But it may be that no number x satisfies the requirements, as in the case $2 - 5$. The system of natural numbers is not closed under the operation of subtraction.

Similarly, children often are taught to divide 6 by 2 by asking "By what must we multiply 2 in order to get 6?" Formalizing this question, we get the following definition.

Definition 2 (*a*). *If a and b are natural numbers, then the symbol $a \div b$ means a natural number x such that $bx = a$, provided such a number x exists. The symbol $a \div b$ is called the* ***quotient*** *a divided by b.*

Definition 2 (*b*). *If there exists a natural number x such that $bx = a$, we say that a is a* ***multiple of*** *b, or a is* ***divisible by*** *b, or b is a* ***factor of*** *a.*

Example. 6 is a multiple of 2 and 2 is a factor of 6 since $2 \cdot 3 = 6$.

Note that we define the new terms "$a \div b$" and "a is a multiple of b" in terms of the old term "multiplication."

The quotient of two natural numbers need not exist. Thus $6 \div 2$ means a natural number x such that $2x = 6$; or $x = 3$, since we discover from the multiplication table that $2 \cdot 3 = 6$. Similarly $5 \div 2$ means a natural number x such that $2x = 5$; but no such natural number exists. Hence $5 \div 2$ is at present a meaningless symbol. The system of natural numbers is not closed under the operation of division.

We shall often use the following postulates of equality.

E_6. *If* $a = b$ *and* $c = d$, *all letters representing natural numbers, then* $a - c = b - d$, *provided these differences exist.* (*If equals are subtracted from equals, the results are equal.*)

E_7. *If* $a = b$ *and* $c = d$, *all letters representing natural numbers, then* $a \div c = b \div d$, *provided these quotients exist.* (*If equals are divided by equals, the results are equal.*)

We might raise the natural question: if there is a number x satisfying the requirements of definition 1, may there not be more than one such number? That is, is $a - b$ a uniquely determined number? There are many instances of symbols whose meanings are not uniquely determined. For example, if A and B are a husband and wife, the symbol "son of A and B" may not have a uniquely determined meaning for there may be more than one person satisfying the requirements of the definition of this symbol. That the numbers $a - b$ and $a \div b$ *are* uniquely determined, if they exist at all, follows at once from the postulates E_6 and E_7 respectively. Therefore the use of the definite article "the" in the expressions "the difference" and "the quotient" is justified.

Only a novice would calculate the value of the expression $(675 + 348) - 675$ by actually adding the numbers in the parentheses and then subtracting, although this is what the expression seems to call for. That this labor is needless is shown by the following theorem.

THEOREM 1. *If* p *and* q *are any natural numbers, then* $(p + q) - p = q$.

Hypothesis. p and q are natural numbers.

Conclusion. $(p + q) - p = q$.

Proof. By A_1, $p + q$ is a natural number. To prove that one natural number $(p + q)$ minus a second (p) equals a third (q), we have only to prove that the second plus the third equals the first, by definition 1(a). But $p + q = (p + q)$ by identity. This completes the proof.

THEOREM 2. *If* p *and* q *are any natural numbers, then* $(pq) \div p = q$.

Hypothesis. p and q are natural numbers.

Conclusion. $(pq) \div p = q$.

Proof. By M_1, (pq) is a natural number. To prove that one natural number (pq) divided by a second (p) equals a third (q), we have only to prove that the second multiplied by the third equals the first, by definition 2(*a*). But $pq = (pq)$ by identity. This completes the proof.

Remark. Subtraction is called the *inverse* of addition and division is called the *inverse* of multiplication, for obvious reasons. Theorems 1 and 2 show that subtraction of x undoes addition of x and division by x undoes multiplication by x.

Remark. Note that in proving theorems 1 and 2 we have made use, essentially, only of definitions 1 and 2. One must guard against the temptation to say, in proving theorem 1 for example, that

$$(p + q) - p = p + q - p = p - p + q = 0 + q = q.$$

For we have no number zero at our command, since the only numbers we know about at present are the natural numbers, and we have never assumed any "commutative law" justifying the second step above. In proving theorems about subtraction, division, greater than, etc., we must not use any intuitive notions, since these terms are defined, not undefined. *We must prove any theorem about a defined term in the light of its definition.*

The existence of the distributive law connecting multiplication with addition suggests that we might have a similar law with subtraction. This is provided by the following theorem.

THEOREM 3. *If a, p and q are natural numbers, and if $p - q$ exists, then $a(p - q) = ap - aq$.*

Hypothesis. a, p, and q are natural numbers; $p - q$ exists.

Conclusion. $a(p - q) = ap - aq$.

Proof. Since $p - q$ exists by hypothesis, we may let $x = p - q$. By definition 1 (*a*), this means that

(1) $$q + x = p.$$

This statement involves p and q as individual terms. What we

want to prove involves aq and ap. Hence it is natural to think of multiplying both sides of (1) by a. Now

(2) $$a = a$$

by identity. Hence, applying E_5 to (1) and (2), we obtain

$$a(q + x) = ap.$$

Now $$aq + ax = ap \text{ (by } D \text{ and substitution).}$$

This means, by definition 1 (a), that

(3) $$ax = ap - aq.$$

But $x = p - q$. Hence, substituting in (3),

$$a(p - q) = ap - aq \qquad \text{(by substitution).}$$

This completes the proof.

THEOREM 4. *If $a > b$ and $b > c$, then $a > c$, all letters representing natural numbers.*

Hypothesis. a, b, c are natural numbers;
$a > b$; $b > c$.

Conclusion. $a > c$.

Proof. By definition 1(b), $a > b$ means there exists a natural number x such that

(4) $$b + x = a.$$

Similarly $b > c$ means that there exists a natural number y such that

(5) $$c + y = b.$$

We have to prove that $a > c$; that is, that there exists a natural number z such that $c + z = a$. Substituting (5) in (4), both being correct by hypothesis, we obtain

$$(c + y) + x = a.$$

By A_3, this becomes $c + (y + x) = a$. But $y + x$ is a natural number by A_1 and this is the "z" we have been looking for. This proves the theorem.

THEOREM 5. *If p and q are natural numbers and if $p - q$ exists, then $(p - q) + q = p$.*

Hypothesis. p and q are natural numbers;
$p - q$ exists.

Conclusion. $(p - q) + q = p$.

Proof. By hypothesis, there is a natural number x such that $q + x = p$. By A_2,

$$x + q = p. \tag{6}$$

But $x = p - q$ by definition. Substituting in (6), we get $(p - q) + q = p$ which was to be proved.

Theorem 3 is an instance of what we shall refer to as the **generalized distributive law** which we shall not prove here but which we shall use in later sections. We shall write *

$$a(b + c - d - e + \cdots + k) = ab + ac - ad - ae + \cdots + ak$$

or

$$(b + c - d - e + \cdots + k)a = ba + ca - da - ea + \cdots + ka$$

whenever these expressions have a meaning. Do not use this generalized distributive law in the exercises below.

EXERCISES

All letters represent natural numbers.

1. Explain what is meant by the statement $10 \div 2 = 5$ in terms of definition 2 (*a*).

2. Explain what is meant by the statement $10 - 2 = 8$ in terms of definition 1 (*a*).

3. Which of the following expressions are meaningless? Explain. (*a*) $8 \div 2$. (*b*) $3 \div 2$. (*c*) $3 - 7$. (*d*) $2 \div 3$. (*e*) $2 \div 8$. (*f*) $7 - 3$.

4. Do the following expressions *always* represent a natural number? (*a*) $a + b$. (*b*) $a - b$. (*c*) ab. (*d*) $a \div b$.

5. If your answer to parts of exercise 4 is negative, what restriction must be placed on a and b in order for these expressions to represent natural numbers?

6. Prove the following, justifying each step by means of a postulate, definition, or previously proved theorem.

(*a*) $(m + n) - n = m$.
(*b*) $(mn) \div n = m$.
(*c*) If $m - n$ exists, then $n + (m - n) = m$.
(*d*) If $m \div n$ exists, then $(m \div n) \cdot n = m$. (Hint: use def. 2 (*a*).)
(*e*) $m \div m = 1$. (Hint: use def. 2 (*a*).)
(*f*) If $p - q$ exists, then $(p - q)a = pa - qa$. (Hint: see Theorem 3.)
(*g*) $(ab + ac) \div a = b + c$. (Hint: use def. 2 (*a*).)
(*h*) $(a + b) \cdot (b + c) \div (a + b) = b + c$. (Hint: use def. 2 (*a*).)
(*i*) If a is a multiple of b and b is a multiple of c, then a is a multiple of c. (Hint: use def. 2(*b*). Compare Theorem 4.)
(*j*) If $a < b$ and $b < c$, then $a < c$.

* In an expression like $b + c - d - e + \cdots + k$ it is understood that the operations indicated are to be performed in the given order, reading from left to right.

(*k*) If a is a factor of b and b is a factor of c, then a is a factor of c.
(*l*) If $a < b$ then $a + c < b + c$.
(*m*) If $a < b$ then $ac < bc$.

16. Fractions, equality, multiplication, and division. Early in history, practical problems of measurement, involving the division of things (such as real estate or quantities of grain) into equal parts, must have led to the invention of fractions. As long as we have only natural numbers at our disposal, division of 4 by 6 is impossible, for there exists no natural number x such that $6x = 4$. In order to make division of one natural number by another always possible without restriction, we invent fractions. To be entirely naive about it, let us notice that when we write a fraction, like 2/3, we are merely writing down a symbol consisting of two natural numbers: 2 and 3. Hence we make the following definition.

DEFINITION 1. *A* ***fraction*** *is a symbol* $\frac{a}{b}$ *(or* a/b*) where* a *and* b *are natural numbers (read "*a *over* b*"). We call* a *the* **numerator** *and* b *the* ***denominator.***

For example, 4/6 and 6/4 are fractions. We wish to apply the symbol 4/6 to represent four of six equal parts of something. Notice that 4/6 does not have anything to do with division at present; it is merely a symbol composed of two natural numbers.

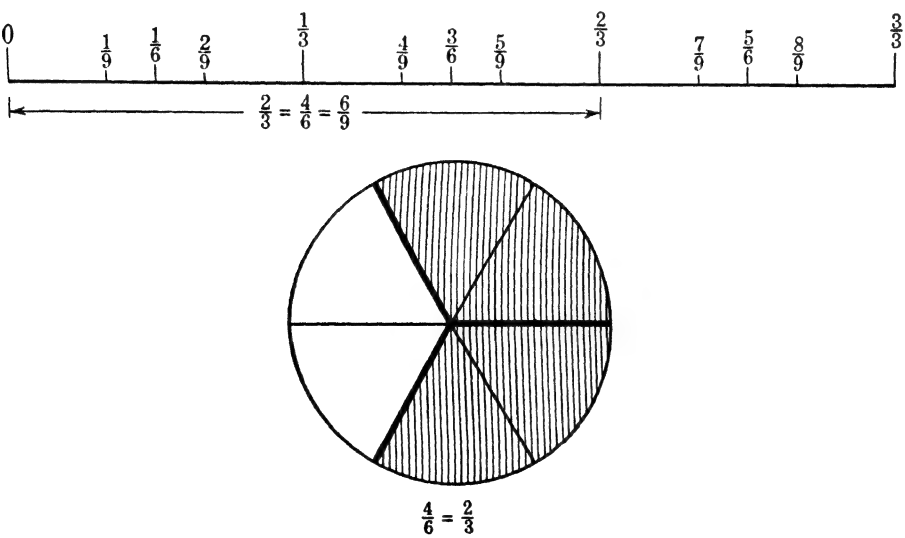


FIG. 17

We might equally well have used some other symbol, like (a, b), for example; thus 4/6 would be written (4, 6). We shall retain the familiar symbolism to avoid confusion. Be careful, however, to read 4/6 as "four over six," not "four divided by six." Notice that we have defined the new term "fraction" in terms of the old term "natural number."

The fractions 4/6, 2/3, 6/9, 8/12 are all different; that is, they are different symbols. But because we have in mind the division of things into equal parts as the application for our abstract fractions, we would like the different fractions 4/6, 2/3, 6/9, 8/12 to be regarded as equal (Fig. 17). Notice that

$$\frac{4}{6} = \frac{2 \cdot 2}{3 \cdot 2}, \quad \frac{6}{9} = \frac{2 \cdot 3}{3 \cdot 3}, \quad \frac{8}{12} = \frac{2 \cdot 4}{3 \cdot 4}.$$

Of course, since fractions are merely symbols created by us, we are free to decide what we shall mean by "equal" fractions. Our practical application to division into equal parts suggests the following preliminary definition.

Definition 2*a*. *If a, b, and x are any natural numbers then*

$$\frac{ax}{bx} = \frac{a}{b}.$$

This means that we may multiply numerator and denominator of a fraction by the same natural number x and obtain an equal fraction. Multiplying numerator and denominator of 2/3 by 2 to get 4/6 simply means concretely that we split each of our 3 equal parts (Fig. 17) into halves and take twice as many of the new parts. Conversely, definition 2a means that we may remove a common factor from the numerator and denominator of a fraction, and obtain thereby an equal fraction. In high school this was called "cancelling." * Thus

$$\frac{4}{6} = \frac{2 \cdot 2}{3 \cdot 2} = \frac{2}{3}.$$

By M_2, it follows from definition $2a$ that we may also write

$$\frac{xa}{xb} = \frac{a}{b}. \tag{1}$$

* It may surprise some students to learn that "cancellation" is not a license to strike out any symbol in both places whenever it appears twice on the same paper or blackboard.

The word denominator is analogous to the word denomination, used in connection with money. For example, two quarters of a dollar may be expressed by the fraction 2/4, the denominator telling the kind of part of the dollar (denomination of the coin) we are talking about while the numerator tells the number of such parts we are considering. To find out whether or not two quarters (2/4) is equal to 5 dimes (5/10) we first express each amount in a common denomination. Thus two quarters may be expressed as 50 cents and 5 dimes may be expressed as 50 cents. We want to regard 2/4 as equal to 5/10 since

$$\frac{2}{4} = \frac{2 \cdot 25}{4 \cdot 25} = \frac{50}{100} \qquad \text{(by Def. 2a)}$$

and

$$\frac{5}{10} = \frac{5 \cdot 10}{10 \cdot 10} = \frac{50}{100} \qquad \text{(by Def. 2a)}$$

and we would of course like to have the statement "things equal to the same thing are equal to each other" remain valid even for fractions.* But the statement $2/4 = 5/10$ can not be obtained from definition 2a since there is no natural number x such that $2x = 5$ and $4x = 10$. Nevertheless we want to be able to say that $2/4 = 5/10$. This suggests the following definition.

DEFINITION 2*b*. *Two fractions shall be* **equal** *if and only if, when they are expressed with the same denominator* (*by means of definition 2a, if necessary*), *they then have the same numerator.*

Applying this definition to two arbitrary fractions a/b and c/d we see that

$$\frac{a}{b} = \frac{ad}{bd} \qquad \text{(by Def. 2a)}$$

and

$$\frac{c}{d} = \frac{bc}{bd} \qquad \text{(by (1))}$$

and hence that our two fractions are equal if and only if the numerators ad and bc are equal. This proves the following theorem.

THEOREM 1. *Two fractions a/b and c/d are equal if and only if $ad = bc$.*

* It was assumed only for natural numbers.

We might have adopted the statement of theorem 1 as our definition of equal fractions. But we prefer to think in terms of definition $2b$ because it is more natural. For example, to decide whether or not $\frac{20}{24}$ and $\frac{30}{36}$ are equal, we have only to express them with a common denominator and then compare the numerators. Thus $\frac{20}{24} = \frac{20 \cdot 3}{24 \cdot 3} = \frac{60}{72}$ while $\frac{30}{36} = \frac{30 \cdot 2}{36 \cdot 2} = \frac{60}{72}$. Hence our two given fractions are equal. Note also that $20 \cdot 36 = 24 \cdot 30$ (see theorem 1).

We could now prove, on the basis of our definitions that statements * analogous to postulates E_1, E_2, E_3 for natural numbers hold for fractions as well. We shall not do this here. In particular, for any fraction we may substitute an equal fraction, in any equation.

When all common factors, other than 1, have been removed from numerator and denominator of a fraction, we say that the fraction has been **reduced to lowest terms** or expressed in its **simplest form.** That is, a fraction is in lowest terms when its numerator and denominator are as small as possible. It could be proved † that each fraction can be reduced to a unique simplest form. For example, 4/6 and 6/9 both have the simplest form 2/3. It could also be proved that two fractions are equal if and only if they have the same simplest form. We shall not prove these statements here.

Definitions $2a$ and $2b$ are motivated by the practical application we have in mind for fractions, namely the division of things into equal parts. If we were going to use the symbol a/b to denote the length and width of a rug it would be foolish to use these definitions; for a rug whose length was 4 and width 6 would be represented by 4/6 while 2/3 would represent a rug whose length is 2 and width 3, and it would be silly to regard 4/6 as equal to

* The treatment of equality of fractions given here is not above logical reproach. For example, one should be able to substitute an equal for an equal in any context whatever. But if in the true statement "4 is the numerator of 4/6" we substitute the "equal" 2/3 for 4/6, we obtain the false statement "4 is the numerator of 2/3." This difficulty can be circumvented but we shall not do so here, because of technical complications.

† With the help of further postulates not stated here. A similar remark applies to many of our statements beginning with "It can be proved that ···." Our set of postulates is neither complete nor irreducible.

2/3. But these fractions (symbols) were created by us for our own use and we make definitions and postulates suitable for the use we have in mind for the symbols.

Our application of fractions to division into equal parts also suggests that $\frac{2}{3}\cdot\frac{4}{5}$ should be $\frac{8}{15}$ since $\frac{2}{3}\cdot\frac{4}{5}$ is interpreted practically to mean two thirds *of* four fifths (Fig. 18). This is suggested by various considerations. If an automobile goes at the rate of

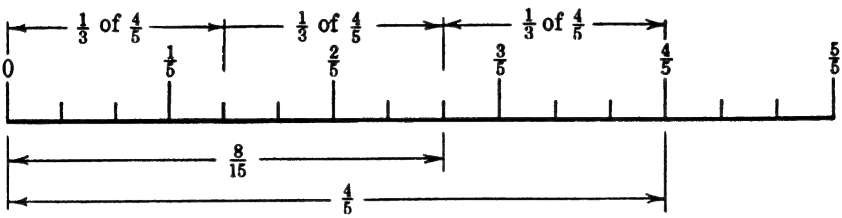


FIG. 18

2 miles per minute for 3 minutes, it covers 6 miles. This suggests the formula $rt = d$, rate times time equals distance, all these quantities being measured in appropriate units.* If a car goes at the rate of $\frac{1}{2}$ mile per minute for $\frac{1}{3}$ of a minute, it covers half of $\frac{1}{3}$ of a mile or $\frac{1}{6}$ of a mile. Hence if we want the same formula to apply to fractional rates and times we are led to define $\frac{1}{2}\cdot\frac{1}{3}$ to be $\frac{1}{6}$ or half *of* $\frac{1}{3}$. Similarly if we sell two dozen eggs at 50 cents per dozen the amount is 100 cents. This suggests the formula $pn = a$, price per item times number equals amount, all these quantities being measured in appropriate units. If we want the same formula to hold for fractional numbers and prices we are led to define $\frac{1}{2}\cdot\frac{1}{3} = \frac{1}{6}$ or half *of* $\frac{1}{3}$ since if we sell $\frac{1}{3}$ of a dozen eggs at the rate of $\frac{1}{2}$ dollar per dozen the amount is $\frac{1}{6}$ of a dollar. This is also suggested by the following geometrical illustration. The area of

* One cannot "multiply" miles per minute by minutes to get miles. One can only multiply numbers by numbers to get numbers. Hence in any such formula it is understood that each letter represents a number (of units of some kind).

a rectangle of length 5 ft. and width 2 ft. is 10 square feet (Fig. 19). Similarly the area of a rectangle $ABCD$ of length 1 ft. and width 1 ft. is 1 square ft. (Fig. 20). This suggests the formula $A = lw$ for the area of a rectangle, all these quantities being measured in appropriate units. Now if we want the area of the rectangle $AEFG$ (Fig. 20) to be given by the same formula we want $\frac{1}{2} \cdot \frac{1}{3}$ to be $\frac{1}{6}$. Notice that the area of $AEFG$ is $\frac{1}{2}$ of the area of $AEHB$ which is $\frac{1}{3}$ sq. ft. Hence we want to interpret $\frac{1}{2} \cdot \frac{1}{3}$ as one half *of* one-third or $\frac{1}{6}$. Similarly $\frac{5}{2} \cdot \frac{4}{3}$ could be interpreted as five halves of 4/3; that is five times four sixths or 20/6. These considerations suggest the following definition.

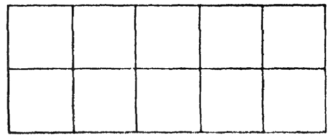


Fig. 19

Definition 3. *If a/b and c/d are any fractions, their product $\frac{a}{b} \cdot \frac{c}{d}$ shall mean the fraction $\frac{ac}{bd}$.*

Note that ac/bd *is* a fraction since ac and bd are both natural numbers, by M_1. Definition 3 says in effect, that the product of two fractions shall be a fraction whose numerator is the product of the two given numerators and whose denominator is the product of the two given denominators.

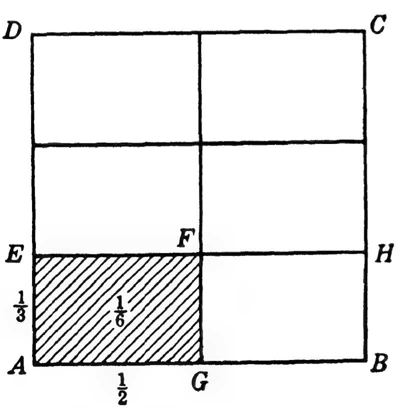


Fig. 20

It can be shown that if a/b or c/d or both are replaced by equal fractions then the resulting product remains equal to ac/bd. A similar remark will hold for the quotient, sum, and difference of two fractions.

It must be emphasized that the paragraph preceding Def. 3 must not be construed as a "proof" that the definition is "so." This paragraph is merely an intuitive explanation of why we *choose* this definition out of all possible definitions.

As in section 15, we may now define division of two fractions as the inverse of multiplication.

DEFINITION 4. *The **quotient** of* $\frac{a}{b} \div \frac{c}{d}$ *shall mean another fraction* $\frac{x}{y}$ *such that* $\frac{c}{d} \cdot \frac{x}{y} = \frac{a}{b}$, *provided such a fraction exists.*

THEOREM 2. *The quotient of two fractions* $\frac{a}{b} \div \frac{c}{d}$ *always exists; in fact,* $\frac{a}{b} \div \frac{c}{d} = \frac{ad}{bc}$.

Proof. We have only to prove that ad/bc satisfies the definition of quotient. That is, we have to show that $\frac{c}{d} \cdot \frac{ad}{bc} = \frac{a}{b}$. But $\frac{c}{d} \cdot \frac{ad}{bc} = \frac{cad}{dbc}$ by definition 3. By M_2, $\frac{cad}{dbc} = \frac{acd}{bcd}$. By definition 2a, $\frac{acd}{bcd} = \frac{a}{b}$. Hence, by substitution, $\frac{c}{d} \cdot \frac{ad}{bc} = \frac{a}{b}$. This completes the proof.

This theorem will be recognized as corresponding to the rule learned by rote in school; namely, to divide one fraction by another, "invert" the second and multiply.

We would like to identify the fraction a/b with the quotient $a \div b$, so that division of one natural number by another would always be possible. But if we did that we would have $3 \div 1 = \frac{3}{1}$ and also $3 \div 1 = 3$ by definition 2 of section 15. But the natural number 3 is not the same as the fraction $3/1$ which is a symbol composed of two natural numbers. Hence we are led to make the following agreement, to prevent such a conflict of meaning.

A fraction $\frac{a}{1}$ *with 1 as denominator shall be written interchangeably with the natural number* a.

This might also have been motivated by our concrete application of fractions to division of things into equal parts. Just as $\frac{3}{4}$ ft. may be interpreted as "divide a foot into 4 equal parts and take 3 of these fourths," we would like to interpret $\frac{3}{1}$ ft. to mean "divide" a foot into one part (that is, leave it undivided) and

take 3 of these "parts." The result of this is of course 3 ft. Hence we would like to regard 3/1 as equal to 3.

THEOREM 3. *If a and b are any two natural numbers then* $a \div b = \dfrac{a}{b}$.

Proof. By our agreement, $a \div b = \dfrac{a}{1} \div \dfrac{b}{1}$. By theorem 2, $\dfrac{a}{1} \div \dfrac{b}{1} = \dfrac{a \cdot 1}{1 \cdot b} = \dfrac{a}{b}$. This completes the proof.

The fraction line may now be used interchangeably with the division sign, although we shall not prove this completely here. The fractions (including the natural numbers which are now identified with fractions whose denominator is 1) have the advantage that *any* two fractions may be divided. We could now prove that the statement "if equals are multiplied or divided by equals, the results are equal" is true for fractions. (This was assumed only for natural numbers, of course.)

We shall use the symbols E_1, E_2, E_3, etc. freely to refer to analogous statements for fractions as well as for natural numbers.

Example. As an example, we give another proof of theorem 1, which is really a theorem and its converse.

(*a*) *If* $\dfrac{a}{b} = \dfrac{c}{d}$ *then* $ad = bc$.

Proof. By hypothesis, $\dfrac{a}{b} = \dfrac{c}{d}$. Multiplying both sides by the common denominator bd we have

$$bd \cdot \frac{a}{b} = bd \cdot \frac{c}{d} \qquad \text{(by } E_5\text{).}$$

Or, $$\frac{bd}{1} \cdot \frac{a}{b} = \frac{bd}{1} \cdot \frac{c}{d} \qquad \text{(by agreement).}$$

Hence, $$\frac{bda}{b} = \frac{bdc}{d} \qquad \text{(by def. 3).}$$

Therefore, $$da = bc. \qquad \text{(Reasons?)}$$
Or, $$ad = bc \qquad \text{(by } M_2\text{).}$$

(*b*) *Conversely, if* $ad = bc$ *then* $\dfrac{a}{b} = \dfrac{c}{d}$.

Proof. By hypothesis, $ad = bc$. Dividing both sides by bd, we have

$$\frac{ad}{bd} = \frac{bc}{bd} \qquad \text{(by } E_7\text{)}.$$

Hence,

$$\frac{a}{b} = \frac{c}{d} \qquad \text{(by def. 2a)}.$$

Notice that when we write $\frac{2}{3} = \frac{2 \cdot 2}{3 \cdot 2} = \frac{4}{6}$ we obtain an equal fraction by definition. This is in harmony with the fact that $\frac{2 \cdot 2}{3 \cdot 2} = \frac{2}{3} \cdot \frac{2}{2}$ (by definition of multiplication of fractions) and therefore $\frac{2}{3} = \frac{2 \cdot 2}{3 \cdot 2} = \frac{2}{3} \cdot \frac{2}{2} = \frac{2}{3} \cdot 1 = \frac{2}{3}$ since any number a times 1 should yield a as the product.

EXERCISES

All letters represent natural numbers.

1. Apply our definitions or theorems to calculate the indicated products and quotients and reduce to simplest form:

(*a*) $\frac{2}{3} \cdot \frac{5}{7}$; (*b*) $\frac{2}{3} \div \frac{5}{7}$; (*c*) $\frac{2}{3} \cdot \frac{6}{4}$; (*d*) $\frac{2}{3} \div \frac{4}{9}$; (*e*) $\frac{2}{3} \div \frac{8}{12}$; (*f*) $\frac{2}{3} \div \frac{16}{12}$.

2. Write a fraction equal to 2/3 having the denominator 21.

3. Write a fraction equal to a/b having the denominator bd.

4. Prove, justifying each step by means of our definitions, theorems, and postulates:

(*a*) $\frac{ab}{cd} \div \frac{b}{d} = \frac{a}{c}$. (*b*) $\frac{2a}{3b} \div \frac{1}{bc} = \frac{2ac}{3}$. (*c*) $\frac{ab + ac}{a} = b + c$.

(*d*) $\frac{ab + ac}{df} \cdot \frac{d}{a} = \frac{b + c}{f}$. (*e*) $\frac{mx + my}{ad} \div \frac{m}{ad} = x + y$.

(*f*) $\frac{ax + ay}{bc + bd} \div \frac{x + y}{c + d} = \frac{a}{b}$.

5. Prove that if $\frac{a}{b} = \frac{c}{d}$ then $\frac{a}{c} = \frac{b}{d}$.

6. Prove that if $\frac{a}{b} = \frac{c}{d}$ then $\frac{b}{a} = \frac{d}{c}$.

7. Prove that if $\frac{a}{b} = \frac{a + c}{b + d}$ then $\frac{a}{b} = \frac{c}{d}$.

8. State and prove the converse of exercise 7.

9. Prove that if $\frac{a+b}{b} = \frac{c+d}{d}$ then $\frac{a}{b} = \frac{c}{d}$.

10. State and prove the converse of exercise 9.

17. Addition and subtraction of fractions. Since fractions are merely symbols, created by us, it is up to us to decide what we shall mean by the sum of two fractions. Here again our choice of definition will be motivated by the practical application we have in mind, namely, the division of things into equal parts.

To add coins of the same denomination we merely add the numbers of coins; for example, two quarters plus one quarter yields three quarters. If two fractions have the same denominator we would naturally like their sum to be a fraction with the same denominator whose numerator is the sum of the two given numerators. For example, we would like $\frac{3}{6} + \frac{2}{6}$ to be $\frac{5}{6}$. This suggests the following definition.

DEFINITION 1. $\frac{a}{b} + \frac{c}{b} = \frac{a+c}{b}$.

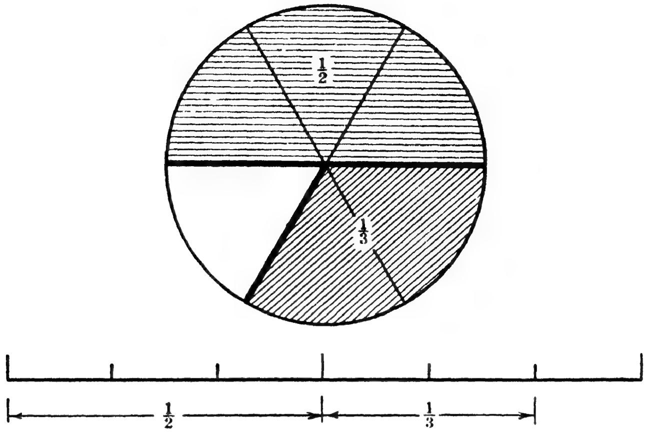
Note that the right member of definition 1 is a fraction since $a + c$ *is* a natural number by A_1.

To add coins of different denominations we first express them in a common denomination, and then add as before. For example, two quarters and three dimes may be added by saying that two quarters is equivalent to 50 cents and three dimes is equivalent to 30 cents; hence the sum is 80 cents. Analogously, it is natural, bearing our applications in mind (Fig. 21), to add 1/2 and 1/3 as follows. By def. 2a, section 16, $\frac{1}{2} = \frac{1 \cdot 3}{2 \cdot 3} = \frac{3}{6}$ and $\frac{1}{3} = \frac{1 \cdot 2}{3 \cdot 2} = \frac{2}{6}$; hence $\frac{1}{2} + \frac{1}{3} = \frac{3}{6} + \frac{2}{6} = \frac{5}{6}$. This leads us to make the following definition.

DEFINITION 2. *The* **sum** *of two fractions with different denominators shall be the fraction obtained by replacing these fractions by equal fractions having a common denominator (by def. 2a, section 16) and then adding according to definition 1.*

It can be proved that the same result is obtained no matter

what common denominator is chosen. One commonly uses the least common denominator.



$\frac{1}{2} + \frac{1}{3} = \frac{5}{6}$

FIG. 21

Example. If $\dfrac{a}{b}$ and $\dfrac{c}{d}$ are any two fractions we have

$$\frac{a}{b} + \frac{c}{d} = \frac{ad}{bd} + \frac{bc}{bd} \qquad \text{(by def. 2a, section 16).}$$

Hence

$$\text{(1)} \qquad \frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \qquad \text{(by def. 1).}$$

Note that ad, bc, and bd are natural numbers by M_1; hence $ad + bc$ is a natural number by A_1. Therefore $\dfrac{ad + bc}{bd}$ *is* a fraction. Instead of going through this process every time we want to add fractions, we might have dispensed with the definitions 1 and 2 altogether and adopted (1) as a general definition. However, it is more natural, and, in practice, it is often more convenient to use our definitions instead of the general definition (1). For example, if we add $\dfrac{1}{2} + \dfrac{3}{8}$ by definition (1) we obtain $\dfrac{1 \cdot 8 + 2 \cdot 3}{2 \cdot 8} = \dfrac{14}{16}$ which has to be reduced to 7/8. But if we add $\dfrac{1}{2} + \dfrac{3}{8}$ by our definitions we say $\dfrac{1}{2} + \dfrac{3}{8} = \dfrac{1 \cdot 4}{2 \cdot 4} + \dfrac{3}{8}$, by def. 2a

section 16, and then $\frac{4}{8} + \frac{3}{8} = \frac{7}{8}$, by definition 1, which doesn't need to be reduced to lowest terms.

In school, the reader may have been taught to add fractions by rote, using blindly memorized mechanical schemes which looked like this

$$\begin{array}{c|c} \frac{1}{3} & 2 \\[1ex] \frac{1}{2} & 3 \\[1ex] \hline \frac{5}{6} & 6 \end{array} \qquad \text{or this} \qquad \begin{array}{c} \overset{2}{\smile}\quad \overset{3}{\smile} \\ \frac{1}{3} + \frac{1}{2} \quad < 6 \\[1ex] \frac{2 \cdot 1 + 3 \cdot 1}{6} = \frac{5}{6} \end{array}$$

or some other scheme which seems to be expressly designed to hide the plain sense of what is going on. If you do not recognize these hieroglyphics, so much the better.

Notice that we have been led to make definitions 1 and 2 because of the application or concrete interpretation we have in mind for fractions, namely problems of measurement involving the division of things into equal parts. Many students attempt to add $\frac{5}{7} + \frac{2}{3}$ by writing $\frac{5+2}{7+3} = \frac{7}{10}$ adding numerators and denominators separately. This is analogous to multiplication where we *do* write $\frac{5}{7} \cdot \frac{2}{3} = \frac{5 \cdot 2}{7 \cdot 3} = \frac{10}{21}$, multiplying numerators and denominators separately. Now the definition $\frac{a}{b} + \frac{c}{d} = \frac{a+c}{b+d}$ which these independent souls persist in using seems a very natural one to write down. It certainly looks simpler, and hence more desirable, than the one we adopted. But it does not fit in with the applications we have in mind. For example, $\frac{1}{2} + \frac{1}{3} = \frac{5}{6}$ fits well with Fig. 21 but $\frac{1}{2} + \frac{1}{3} = \frac{2}{5}$ would not. However, there is no logical reason why we could not make the definition $\frac{a}{b} + \frac{c}{d} = \frac{a+c}{b+d}$ if we wanted to. We would, of course, get an algebra which looks quite different from the one we are used to and we would, of course, be unable to apply it to the division of things into equal

parts. But this new strange algebra might have *other* applications. For example, if we want to use the fraction to represent the number of games a baseball team has won as compared with the number it has played, we might do it as follows. If a team has played a series of 7 games and won 5, we would express this fact by the fraction 5/7. Suppose the team now plays a series of 3 games and wins 2 of these. The result of this series is expressed by the fraction 2/3. Now if we want the total result of both series to be obtained by adding the two fractions together, it is natural to use the definition $\frac{a}{b} + \frac{c}{d} = \frac{a + c}{b + d}$ and write $\frac{5}{7} + \frac{2}{3} = \frac{7}{10}$, thus expressing the fact that the team has won 7 games out of 10. This brings into the foreground again the idea that we can choose our postulates and definitions in any self-consistent way and get different abstract mathematical sciences which may have different concrete applications. In Chapter VI we shall in fact discuss algebras different from the one we are used to and which do have important applications.

We now define subtraction of fractions as the inverse of addition.

DEFINITION 3. *If* $\frac{a}{b}$ *and* $\frac{c}{d}$ *are any two fractions, the* ***difference*** $\frac{a}{b} - \frac{c}{d}$ *shall mean the fraction* $\frac{x}{y}$ *such that* $\frac{c}{d} + \frac{x}{y} = \frac{a}{b}$, *provided such a fraction exists. If such a fraction does exist, we shall say that* $\frac{a}{b} > \frac{c}{d}$ *or* $\frac{c}{d} < \frac{a}{b}$.

Expressing the fractions $\frac{a}{b}$ and $\frac{c}{d}$ with a common denominator we have $\frac{ad}{bd}$ and $\frac{bc}{bd}$. It would seem natural to subtract one from the other by subtracting numerators and retaining the common denominator, just as one would do with coins of the same denomination. For example, three quarters minus one quarter is two quarters. This would yield the fraction $\frac{ad - bc}{bd}$ as the difference $\frac{ad}{bd} - \frac{bc}{bd}$ provided bc can be subtracted from ad, that is, if $ad > bc$. This result is established by the following theorem.

THEOREM 1. *If* $ad > bc$, *then* $\frac{a}{b} - \frac{c}{d} = \frac{ad - bc}{bd}$.

Proof. Note that if ad were not greater than bc then $ad - bc$ would not be a natural number and $\frac{ad - bc}{bd}$ would not be a fraction at all. We have to prove that $\frac{ad - bc}{bd}$ satisfies definition 3; that is, that $\frac{c}{d} + \frac{ad - bc}{bd} = \frac{a}{b}$. Now

$$\begin{aligned}
\frac{c}{d} + \frac{ad - bc}{bd} &= \frac{bc}{bd} + \frac{ad - bc}{bd} && \text{(Def. 2a, section 16)}\\
&= \frac{bc + (ad - bc)}{bd} && \text{(Def. 1)}\\
&= \frac{ad}{bd} && \text{(A_2 and Theorem 5, section 15)}\\
&= \frac{a}{b} && \text{(Def. 2a, section 16).}
\end{aligned}$$

This completes the proof.

The converse of this theorem can also be proved; in fact, it can be proved that the three statements "$\frac{a}{b} - \frac{c}{d}$ exists," "$\frac{a}{b} > \frac{c}{d}$," and "$ad > bc$" are equivalent. (We shall not prove this here.) This last statement means that *one fraction is greater than another if and only if its numerator is greater than the numerator of the other when they have both been expressed with a common denominator.* Thus to decide whether or not 21/24 is greater than 31/36 we may write $\frac{21}{24} = \frac{21 \cdot 3}{24 \cdot 3} = \frac{63}{72}$ and $\frac{31}{36} = \frac{31 \cdot 2}{36 \cdot 2} = \frac{62}{72}$; hence $\frac{21}{24} > \frac{31}{36}$. We might also have decided this from the fact that $21 \cdot 36 = 756 > 744 = 24 \cdot 31$ since $\frac{21}{24} = \frac{21 \cdot 36}{24 \cdot 36} = \frac{756}{864}$ while $\frac{31}{36} = \frac{31 \cdot 24}{36 \cdot 24} = \frac{744}{864}$. In general $\frac{a}{b} = \frac{ad}{bd}$ while $\frac{c}{d} = \frac{bc}{bd}$; hence we can verify that $\frac{a}{b} > \frac{c}{d}$ by merely observing that $ad > bc$.

In practice, it is unnecessary to memorize theorem 1. To subtract fractions, simply express them with a common denomina-

tor and subtract the numerators, as follows: $\frac{1}{2} - \frac{1}{3} = \frac{3}{6} - \frac{2}{6} = \frac{3-2}{6} = \frac{1}{6}.$

We could now prove easily that the statements "if equals are added to, or subtracted from, equals, the results are equal" is true for fractions, but we shall not do so here. (These statements were assumed only for natural numbers.) We shall use freely the symbols E_6 and E_7 to refer to these analogous statements for fractions as well as for natural numbers.

Note that if we express natural numbers a and c as fractions $a/1$, $c/1$, respectively, then the operations defined for fractions yield the same results as the former operations for natural numbers, wherever the latter have a sense. For example, $\frac{a}{1} \cdot \frac{c}{1} = \frac{ac}{1}$ and $\frac{a}{1} + \frac{c}{1} = \frac{a+c}{1}$ which is what we would get from our operations for the natural numbers a and c if we did not write them as fractions. This is essentially why "identifying" a with $a/1$ causes no trouble.

It can be shown that if a and b are any two unequal fractions, either $a < b$ or $a > b$, but not both.

EXERCISES

1. Apply our definitions, theorems and postulates to calculate the indicated sums and differences, wherever possible, and reduce to lowest terms:

(a) $\frac{2}{3} + \frac{3}{2}.$ (b) $\frac{3}{2} - \frac{2}{3}.$ (c) $\frac{2}{3} - \frac{3}{2}.$ (d) $\frac{2}{3} - \frac{2}{3}.$

(e) $\left(\frac{2}{3} + \frac{3}{4}\right) + \frac{7}{12}.$ (f) $\frac{2}{3} + \left(\frac{3}{4} + \frac{7}{12}\right).$

2. If a and b are any two *fractions*, does there always exist a fraction

(a) $a + b$; (b) $a - b$; (c) ab; (d) $a \div b$;

and, if not, what restrictions must be placed on a and b in order to insure the existence of the result?

3. Prove, justifying each step by a definition, postulate, or theorem, all letters representing natural numbers:

(a) $\frac{1}{x} + \frac{1}{2} = \frac{2+x}{2x}.$ (b) $\frac{y}{2x} + \frac{1}{2} = \frac{y+x}{2x}.$ (c) $x + \frac{1}{y} = \frac{xy+1}{y}.$

(d) $\left(\frac{a}{b} + 3\right) \div \left(\frac{a}{bc} + \frac{3}{c}\right) = c.$ (e) $\left(\frac{1}{a} + \frac{1}{b}\right) \div \left(\frac{2}{a} + \frac{2}{b}\right) = \frac{1}{2}.$

(*f*) $\left(\frac{ad}{bx} + \frac{2d}{x}\right) \div \left(\frac{a}{bc} + \frac{2}{c}\right) = \frac{cd}{x}.$

(*g*) $\left(\frac{a}{b} + \frac{c}{d}\right) + \frac{e}{f} = \frac{adf + bcf + bde}{bdf}.$

(*h*) $\frac{a}{b}\left(\frac{c}{d} + \frac{e}{f}\right) = \frac{acf + ade}{bdf}.$

(*i*) $\left(\frac{1}{a} + \frac{1}{b}\right) \div \left(\frac{1}{a} - \frac{1}{b}\right) = \frac{b + a}{b - a}$, assuming $a < b$.

4. Is 34/51 equal to, greater than, or less than 38/57?

5. Is 34/51 equal to, greater than, or less than 48/69?

18. Some properties of fractions. We have seen in connection with natural numbers that many of the familiar manipulations can be justified by the associative, commutative, and distributive laws. These laws were assumed for natural numbers only. We shall now *prove* that fractions obey them. As a result we shall be able to manipulate fractions in the usual ways.

THEOREM 1. *(Compare A_1). Given any pair of fractions a/b and c/d, in the stated order, there is a third fraction called their sum.* **(Law of closure for addition of fractions.)**

Proof. By definition 2, section 17.

THEOREM 2. *(Compare A_2.) If x/y and u/v are any fractions then*

$$\frac{x}{y} + \frac{u}{v} = \frac{u}{v} + \frac{x}{y}.$$

(Commutative law for addition of fractions.)

Proof. By definitions 1 and 2, section 17, $\frac{x}{y} + \frac{u}{v} = \frac{xv}{yv} + \frac{yu}{yv} = \frac{xv + yu}{yv}$ and $\frac{u}{v} + \frac{x}{y} = \frac{uy}{vy} + \frac{vx}{vy} = \frac{uy + vx}{vy}$. Now $yv = vy$, $xv = vx$ and $yu = uy$ by M_2. Thus $xv + yu = vx + uy$. Finally $vx + uy = uy + vx$ by A_2. The theorem follows by E_3.

THEOREM 3. *(Compare A_3). If x/y, u/v, m/n are any fractions,*

$$\left(\frac{x}{y} + \frac{u}{v}\right) + \frac{m}{n} = \frac{x}{y} + \left(\frac{u}{v} + \frac{m}{n}\right).$$

(Associative law for addition of fractions.)

Proof.

$$\left(\frac{x}{y}+\frac{u}{v}\right)+\frac{m}{n}=\left(\frac{xv}{yv}+\frac{yu}{yv}\right)+\frac{m}{n} \qquad \text{(def. 1, section 17)}$$

$$=\frac{xv+yu}{yv}+\frac{m}{n} \qquad \text{(def. 2, section 17)}$$

$$=\frac{(xv+yu)n}{yvn}+\frac{yvm}{yvn} \qquad \text{(def. 1, section 17)}$$

$$=\frac{(xv+yu)n+yvm}{yvn} \qquad \text{(def. 2, section 17)}$$

(1) $$=\frac{xvn+yun+yvm}{yvn}$$ (by the generalized distributive law for natural numbers).

Similarly,

$$\frac{x}{y}+\left(\frac{u}{v}+\frac{m}{n}\right)=\frac{x}{y}+\frac{un+vm}{vn} \qquad \text{(Reason?)}$$

$$=\frac{xvn+y(un+vm)}{yvn} \qquad \text{(Reason?)}$$

(2) $$=\frac{xvn+yun+yvm}{yvn} \qquad \text{(Reason?)}$$

But the right members of (1) and (2) are the same. Hence, the left members are equal since things equal to the same thing are equal to each other. This completes the proof.

THEOREM 4. (*Compare* M_1). *If a/b and c/d are any fractions, given in the stated order, there exists a third fraction called their product.* **(Law of closure for multiplication of fractions.)**

Proof. By definition 3, section 16.

THEOREM 5. (*Compare* M_2). *If x/y and u/v are any fractions,*

$$\frac{x}{y}\cdot\frac{u}{v}=\frac{u}{v}\cdot\frac{x}{y}.$$

(Commutative law for the multiplication of fractions.)

Proof. Left to the reader.

THEOREM 6. (*Compare* M_3). *If x/y, u/v, m/n are any fractions*

$$\left(\frac{x}{y}\cdot\frac{u}{v}\right)\frac{m}{n}=\frac{x}{y}\left(\frac{u}{v}\cdot\frac{m}{n}\right).$$

(Associative law for the multiplication of fractions.)

Proof. Left to the reader.

THEOREM 7. (*Compare D*). *If x/y, u/v, m/n are any fractions, then*

$$\frac{x}{y}\left(\frac{u}{v}+\frac{m}{n}\right)=\frac{x}{y}\cdot\frac{u}{v}+\frac{x}{y}\cdot\frac{m}{n}.$$

(Distributive law for fractions.)

Proof.

$$\frac{x}{y}\left(\frac{u}{v}+\frac{m}{n}\right)=\frac{x}{y}\left(\frac{un}{vn}+\frac{vm}{vn}\right)=\frac{x}{y}\left(\frac{un+vm}{vn}\right) \qquad \text{(definition 2, section 17)}$$

$$=\frac{x(un+vm)}{yvn} \qquad \text{(definition 3, section 16)}$$

(3) $$=\frac{xun+xvm}{yvn} \qquad (D).$$

Now,

$$\frac{x}{y}\cdot\frac{u}{v}+\frac{x}{y}\cdot\frac{m}{n}=\frac{xu}{yv}+\frac{xm}{yn} \qquad \text{(definition 3, section 16)}$$

$$=\frac{xun}{yvn}+\frac{xmv}{ynv} \qquad \text{(definition 2a, section 16)}$$

$$=\frac{xun}{yvn}+\frac{xvm}{yvn} \qquad \text{(Reason?)}$$

(4) $$=\frac{xun+xvm}{yvn} \qquad \text{(definition 1, section 17).}$$

But the right members of (3) and (4) are equal. Hence the left members are equal since things equal to the same thing are equal to each other. This completes the proof.

We may use these theorems to good advantage in simplifying expressions involving fractions.

Example. Prove that $\frac{x}{y}\left(\frac{cy}{x}+\frac{my}{nx}\right)=\frac{m}{n}+c$. By the distributive law for fractions,

$$\frac{x}{y}\left(\frac{cy}{x} + \frac{my}{nx}\right) = \frac{x}{y} \cdot \frac{cy}{x} + \frac{x}{y} \cdot \frac{my}{nx}$$

$$= \frac{xcy}{yx} + \frac{xmy}{ynx} \qquad \text{(definition 3, section 16)}$$

$$= \frac{c}{1} + \frac{m}{n} \qquad \text{(definition 2a, section 16)}$$

$$= \frac{m}{n} + \frac{c}{1} \qquad \text{(commutative law for addition of fractions)}$$

$$= \frac{m}{n} + c \qquad \text{(by agreement).}$$

The ancient Greeks were acquainted with fractions but regarded a fraction as a relation between two numbers. From the modern point of view, a fraction is considered as a single number *because it does have many important properties in common with the natural numbers, such as the associative, commutative, and distributive laws.*

We have proved that fractions have the properties enunciated for natural numbers in postulates A_1, A_2, A_3, M_1, M_2, M_3, and D. An important consequence of this is that fractions must therefore have the properties enunciated for natural numbers in all the theorems which follow from these postulates alone. That is, any theorem about natural numbers whose proof depends only on the seven postulates mentioned can be converted into a correct theorem about fractions by merely changing the term "natural number" into the term "fraction." This is so because, since the postulates are true of fractions, all their consequences must be too. Another way to put this is that if we based an abstract mathematical science on the seven postulates mentioned, writing mumbo instead of natural number, then interpreting the undefined term mumbo as meaning natural number would yield a concrete interpretation of this mathematical science; while interpreting the undefined term mumbo as meaning fraction would yield a different concrete interpretation of the same abstract mathematical science.

By this time the student is beginning to see how a great many of the manipulations of symbols which he learned by rote in school are really logical consequences of a few simple postulates and definitions which in turn are suggested by experience. Now

it goes without saying that mechanical manipulations are quicker for the technician who does not wish to waste time, and there is no harm in using short cuts and trick devices provided their use is preceded by a clear understanding of the reasoning behind them. However, if mechanical devices *replace* reasoning, the purposes and ideals of mathematics are forgotten and we are merely learning a trade by memory, much as some sea-captains navigate without understanding trigonometry, upon which navigation is based. In any case, *our* purpose is not technical facility but rather understanding of the logical structure of algebra.

EXERCISES

1. Simplify, justifying each step by means of a definition, theorem, or postulate, all letters representing natural numbers.

(*a*) $3\left(\frac{1}{3}+4\right)$. (*b*) $\frac{1}{2}\cdot\frac{5}{6}+\frac{1}{2}\cdot\frac{7}{6}$. (*c*) $\frac{3+6x}{3}$. (*d*) $\frac{2+5}{2+10}$.

(*e*) $\dfrac{\frac{1}{2}+\frac{1}{3}}{\frac{1}{2}-\frac{1}{3}}$. (*f*) $\dfrac{\frac{a}{b}+\frac{c}{d}}{\frac{a}{b}-\frac{c}{d}}$, assuming that $ad > bc$.

2. Prove, justifying each step by means of a definition, theorem, or postulate, all letters representing natural numbers:

(*a*) $\left(\frac{p}{q}+\frac{m}{n}\right)\frac{a}{b}=\frac{am}{bn}+\frac{ap}{bq}$. (*b*) $\frac{a}{b}\left(\frac{bc}{ad}+\frac{be}{a}\right)=\frac{c}{d}+e$.

(*c*) $\frac{6}{5b}\left(\frac{10bc}{3}+\frac{db}{a}\right)=4c+\frac{6d}{5a}$. (*d*) $\frac{1}{x}\left(\frac{x}{y}+x\right)=\frac{1}{y}+1$.

3. Prove, using definition 3 of section 17, and theorems of the present section: if $\frac{a}{b}>\frac{c}{d}$ and $\frac{c}{d}>\frac{e}{f}$ then $\frac{a}{b}>\frac{e}{f}$.

4. Prove that if $\frac{a}{b}>\frac{c}{d}$ then $\frac{a}{b}+\frac{x}{y}>\frac{c}{d}+\frac{x}{y}$.

19. Relations. Exercise 3, section 18, and theorem 4, section 15, suggest that the relation "is greater than" has a very familiar property. We digress to discuss briefly some important properties of *relations* in general. The idea of "relation" is an easy one to grasp. We shall indicate the meaning of the word by some illustrative examples. For example, "being father of," "being an ancestor of," "being taller than," "being equal to," "being greater than," are *relations* in which things of various kinds may

stand to one another. The first three relations mentioned above may hold between people, while the last two may hold between numbers. We shall list three important properties that relations sometimes have.

Consider an arbitrary relation R and let x, y, z, etc. represent things which may stand in this relation to each other.

It *may* happen that for any object x it *must* be true that x has the relation R to itself. If this is the case, the relation R is called **reflexive.** For example, the relation "is equal to," for numbers, is a reflexive relation, since $a = a$ (see E_1). But "is the father of," for people, is definitely not reflexive since no person x is the father of himself.

It *may* happen that whenever x has the relation R to y then y *must* also have the relation R to x. If this is the case, the relation R is called **symmetric.** Thus "is equal to," for numbers, is a symmetric relation since if $a = b$ then $b = a$ (see E_2). But "is the father of," for people, is not symmetric since if x is the father of y it is not true that y is the father of x.

It *may* happen that if x has the relation R to y and y has the relation R to z then x *must* have the relation R to z. If this is the case the relation R is called **transitive.** The relation "is equal to," for numbers, is transitive, since if $a = b$ and $b = c$ then $a = c$ (see E_3). But "is the father of," for people, is not transitive, since if x is the father of y and y is the father of z, it is not true that x is the father of z.

In exercise 3, section 18, we proved that the relation "is greater than," for fractions, is transitive. In theorem 4, section 15, we proved that the relation "is greater than," for natural numbers, is transitive. If you found it hard to see why we bothered to prove these statements, it is doubtless because you have always assumed subconsciously that the relation "greater than" is transitive.

Example 1. Consider the relation "implies," which may hold between propositions p, q, r, etc. This relation is clearly reflexive since p implies p; that is, if p is true then p is true. It is not symmetric, since p implies q may be correct while q implies p is not; that is, a proposition may be correct while its converse is not. It is transitive since if p implies q and q implies r, then p implies r;

in fact, this is an important principle of logic which we use continually in our proofs without the slightest hesitation.

Example 2. Consider the relation "was born in the same town as," which may hold among people. This is clearly reflexive, symmetric, and transitive.

EXERCISES

Which of the adjectives "reflexive," "symmetric," "transitive" is applicable to the following relations?

1. "Is the mother of," for people.
2. "Is an ancestor of," for people.
3. "Is the spouse of," for people.
4. "Is in love with," for people.
5. "Is taller than," for people.
6. "Is less than," for numbers.
7. "Is the husband of," for people.
8. "Is west of," for places in America.
9. "Is west of," for places anywhere on Earth except the poles.
10. "Is a multiple of," for natural numbers.
11. "Is a factor of," for natural numbers.
12. "Has the same length as," for line-segments.
13. "Is perpendicular to," for lines in a plane.
14. "Is not equal to," for numbers.
15. "Is congruent to," for triangles.
16. "Is similar to," for triangles.
17. "Lives within a mile of," for people.

Since "natural number" has had a concrete meaning for us, our study of numbers has not constituted an abstract mathematical science (although it will suggest an abstract treatment of algebra to be discussed briefly in section 28). In any case, neither algebra nor geometry provides a *simple* example of an abstract mathematical science since each rests on a large number of postulates. On the other hand, the examples 1 and 2 of section 8 are simple but trivial. The following is a good example of an abstract mathematical science with several different concrete interpretations; it is neither complicated nor unimportant.

Example. Consider a collection of objects, of unspecified nature, and an undefined relation, of unspecified nature, which may

hold between pairs of these objects. The objects in the collection will be denoted by small letters, like a, b, c, and we shall write $a\ R\ b$ to mean that a has the relation (mentioned above) to b. Assume the following postulates.

POSTULATE 1. *If $a\ R\ b$ then a is different from b.*

POSTULATE 2. *If $a\ R\ b$ and if $b\ R\ c$ then $a\ R\ c$.*

From these postulates we deduce the following theorem.

THEOREM 1. *If $a\ R\ b$ is true then $b\ R\ a$ is false.*

Proof. By hypothesis, $a\ R\ b$ is true. Suppose that $b\ R\ a$ were also true. Then from postulate 2, we would have $a\ R\ a$, because $a\ R\ b$ and $b\ R\ a$ would imply $a\ R\ a$. But by postulate 1, $a\ R\ a$ implies that a is different from a, which is absurd. Hence the supposition that $b\ R\ a$ is true must be false since it leads to a false conclusion. That is, $b\ R\ a$ is false.

We now introduce a defined term.

DEFINITION. *If $a\ R\ b$ and $b\ R\ c$ then b is said to be between a and c.*

THEOREM 2. *If b is between a and c and if c is between b and d, then c is between a and d.*

Proof. The hypothesis "b is between a and c" means that $a\ R\ b$ and $b\ R\ c$. By postulate 2, this implies that $a\ R\ c$. The hypothesis "c is between b and d" implies that $b\ R\ c$ and $c\ R\ d$. But $a\ R\ c$ and $c\ R\ d$ means, by definition, that c is between a and d.

Here we have the beginnings of an abstract mathematical science with undefined terms, postulates, defined terms and theorems. This abstract mathematical science has many important concrete interpretations, some of which we shall now list.

First Interpretation. Let the "objects" be natural numbers and let $a\ R\ b$ mean "a is less than b."

Second Interpretation. Let the "objects" be fractions and let $a\ R\ b$ mean "a is less than b."

Third Interpretation. Let the "objects" be natural numbers and let $a\ R\ b$ mean "a is greater than b."

Fourth Interpretation. Let the "objects" be points on a line and let $a\ R\ b$ mean "a is to the left of b."

Fifth Interpretation. Let the "objects" be instants of time and let $a\ R\ b$ mean "a is before b."

Sixth Interpretation. Let the "objects" be all the natural numbers, and let $a\ R\ b$ mean "a is a proper factor of b." By a **proper factor** of a natural number n is meant any factor except 1 and the number n itself. (Every natural number n has the factors 1 and n, sometimes called its "trivial factors.")

Seventh Interpretation. Let the "objects" be people, and let $a\ R\ b$ mean "a is an ancestor of b."

Eighth Interpretation. Let the "objects" be people, and let $a\ R\ b$ mean "a is taller than b."

Exercise. Verify intuitively that each of these concrete interpretations converts our postulates into true statements, and therefore converts our theorems into true statements as well. Restate postulates, definitions and theorems in terms of each interpretation, in succession.

REFERENCES

E. T. Bell, *The Queen of the Sciences*, G. E. Stechert, N. Y., 1938.

A. A. Bennett and C. A. Baylis, *Formal Logic*, (Chapter III, sections 3.6 and 3.7, Relations), Prentice-Hall, N. Y., 1939.

M. R. Cohen and E. Nagel, *Introduction to Logic and Scientific Method* (Chapter 6, section 2. The Formal Properties of Relations), Harcourt Brace, N. Y., 1934.

T. Dantzig, *Number, The Language of Science*, Macmillan, N. Y., 1939.

A. Dresden, *Invitation to Mathematics*, Holt, N. Y., 1936.

H. B. Fine, *College Algebra*, Part I, Ginn, Boston, 1904.

H. B. Fine, *The Number System of Algebra*, Leach, Shewell, Sanborn, Boston, 1890.

B. Russell, *Introduction to Mathematical Philosophy*, (Chapter 5, "Kinds of Relations"), George Allen & Unwin, London, 1919.

A. N. Whitehead, *Introduction to Mathematics*, Holt, N. Y., 1911.

J. W. Young, *Fundamental Concepts of Algebra and Geometry*, Macmillan, N. Y., 1911.

W. W. R. Ball, *A Short Account of the History of Mathematics*, Macmillan, London, 1908.

F. Cajori, *A History of Elementary Mathematics*, Macmillan, N. Y., 1917.

F. Cajori, *A History of Mathematics*, Macmillan, N. Y., 1919.

L. L. Conant, *The Number Concept*, Macmillan, N. Y., 1896.

K. Fink, *A Brief History of Mathematics*, Open Court, Chicago, 1910.

V. Sanford, *A Short Account of the History of Mathematics*, Houghton Mifflin, Boston, 1930.

D. E. Smith, *History of Mathematics*, Ginn, Boston, 1923, 1925.

Chapter IV

FURTHER EVOLUTION OF THE NUMBER SYSTEM

20. Introduction. In the last chapter we discussed in considerable detail the logical development of the simplest numbers, that is, the natural numbers and the fractions. These numbers satisfied the practical needs of man for many thousands of years and were the only ones dealt with until comparatively modern times. In this chapter we shall discuss still other types of numbers such as negative, irrational, and imaginary numbers. We shall make no attempt to carry through many strictly logical proofs in connection with these more sophisticated kinds of numbers, since such a task would be too difficult. We shall, however, indicate why it was natural for these numbers to be invented and we shall try to dispel some of the misconceptions which you may have about them.

21. Directed numbers. Among the fractions, it is possible to add, multiply, and divide freely, since the sum, product, and quotient of any two fractions exists within the system of fractions (that is, is a fraction). But it is still impossible to subtract a larger fraction from a smaller one, or even a fraction from itself. In other words, the system of fractions is closed under the operations of addition, multiplication, and division, but not under subtraction. We shall remedy this situation by extending the idea of number in a fundamental way. Corresponding to each number already in existence, such as 3, we invent two new symbols or marks, such as $+3$ and -3. Corresponding to the number $\frac{1}{2}$ we invent two new symbols $+\frac{1}{2}$ and $-\frac{1}{2}$. The symbols preceded by a plus sign are called **positive numbers** and those preceded by a minus sign are called **negative numbers.** We also invent a new symbol 0, called **zero.** The symbols 0, $+1$, $+2$, $+3$, and so on,

and -1, -2, -3, and so on, are called **integers,** or **whole numbers.** (The Latin word "integer" means whole.) We speak of positive and negative integers, and of positive and negative fractions. Zero is neither positive nor negative.

All these new symbols are called **directed numbers** or **signed numbers.** They are so-called because they may be interpreted geometrically in terms of direction. This is done by choosing a line, a suitable unit of length, and a starting point on the line which we shall call the origin. To the origin we attach the num-

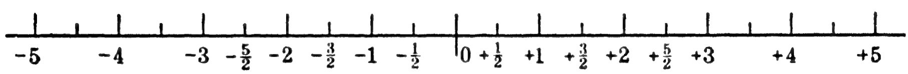


Fig. 22

ber zero. Marking off equal intervals in both directions (Fig. 22), we attach positive integers to marked points in one direction and negative integers in the other. The other positive and negative fractions are also attached to points on the line in the obvious way. It is customary to use the right hand side of the line for the positive numbers.

Notice that the plus and minus signs in the symbols $+3$ and -3 are not intended to indicate the operations of addition and subtraction, but are merely marks to distinguish one direction from another. We might well have denoted $+3$ and -3 by $_R3$ and $_L3$, respectively, to indicate the right-hand 3 and the left-hand 3. However, we shall adhere to the usual notation to avoid confusing you. The numbers studied in the preceding chapter will be called **unsigned numbers** to distinguish them from our new signed or directed numbers.

The idea of a negative number struggled for recognition for centuries and was received with great reluctance as late as the early years of the 17th century, even by mathematicians. Negative numbers were often called false or fictitious numbers. Now it is true that no one ever saw -3 books on a table. But the whole point is that directed numbers are no longer to be interpreted as quantity or magnitude alone. One way of interpreting them is as quantity or magnitude together with direction. To try to think of -3 as referring to taking away 3 books from an empty table is to miss the point completely. But since these new things involve a new idea, namely direction, why do we call them

numbers? Simply because, when we have made the appropriate definitions of addition and multiplication, they will have many of the characteristics of our former numbers, such as the commutative, associative, and distributive laws. In fact, we will be careful to make our definitions of addition and multiplication in such a way as to insure the validity of these laws. As for applications, you are well aware that they are convenient for such things as East and West, up and down, profit and loss, past and future, temperature above and below zero, a push in one direction and a push in the opposite direction, etc.

In fact, to question whether a word, like "number," *can be permitted* to have a certain meaning is foolish, because the meaning of any word or symbol is merely a matter of agreement. All that one may reasonably demand is that the writer inform us what he *intends* a given symbol or word to mean.* This point was encountered by Alice in Lewis Carroll's *Through the Looking Glass* when Humpty Dumpty said:

" 'When *I* use a word, it means just what I choose it to mean—neither more or less.'

" 'The question is,' said Alice, 'whether you *can* make words mean so many different things.'

" 'The question is,' said Humpty Dumpty, 'which is to be the master—that's all.' "

Since these directed numbers were created by us, it is up to us to decide what we shall mean by the sum, product, etc., of two directed numbers. Our definition of the sum of two directed numbers will be motivated by Fig. 22.

Suppose we begin by standing at the origin facing the positive direction, and interpret each successive directed number as a marching order. The directed number $+ 2$ will mean "march forward (that is, to the right) 2 units," while the directed number $- 2$ will mean "march backward (that is, to the left) 2 units." The plus sign for addition will mean "and then." Zero will be interpreted as "do not march at all." The number attached to the point at which we arrive at the end is called the *sum* of the given directed numbers.

* In discussions of social problems, this is, unfortunately, seldom done. The futility of many debates is due to the *undeclared* use of different meanings for the same word.

Example 1. $(+3)+(+2)$ means walk forward 3 units and then walk forward two units. We arrive at the point marked $+5$. Hence $(+3)+(+2)=+5$. Note that the plus signs within the parentheses on the left indicate positive numbers while the plus sign between the two sets of parentheses indicates the operation of addition. It would be preferable to avoid this ambiguity by reserving the plus sign for the operation of addition and to write ${}_R3 + {}_R2 = {}_R5$, but we shall not do this because of the weight of habit and tradition.

Example 2. $(+5)+(-2)$ means walk 5 units forward and then 2 units backward, arriving at the point marked $+3$. Hence, $(+5)+(-2)=+3$, or ${}_R5 + {}_L2 = {}_R3$.

Example 3. $(+2)+(-5)$ means walk forward 2 units and then 5 units backward. Hence $(+2)+(-5)=-3$, or ${}_R2 + {}_L5 = {}_L3$.

Example 4. $(+3)+(-3)$ means walk forward 3 units and then backward 3 units. Hence $(+3)+(-3)=0$, or ${}_R3 + {}_L3 = 0$.

Example 5. $(-3)+(-2)$ means walk backward 3 units and then backward 2 units. Hence $(-3)+(-2)=-5$, or ${}_L3 + {}_L2 = {}_L5$.

Example 6. Clearly any number plus zero will be the given number again, since zero means "do not march at all."

This intuitive idea of walking back and forth on the line of Fig. 22, is expressed formally in the following definition of addition of directed numbers. The formal definition seems complicated only because of the necessity for treating exhaustively all possible cases. For example, we would like $(+5)+(-2)$ to be $+3$; this can be written formally as $(+5)+(-2)=+(5-2)$ where the $5-2$ in the parenthesis on the right means subtraction of the *unsigned* numbers 5 and 2. This subtraction has a sense since $5 > 2$. But we could not write $(+2)+(-5)$ as $+(2-5)$ because the difference $2-5$ of the *unsigned* numbers 2 and 5 does not exist since $2 < 5$. Hence in this case we write $(+2)+(-5) = -(5-2)$ or -3 since the difference $5-2$ of the unsigned numbers 5 and 2 does have a present meaning. The several cases in the following definition are numbered so as to correspond to the

numbered examples above. Note that the definition of the new term "sum of two directed numbers" is given in terms of old terms, namely operations on unsigned numbers.

DEFINITION 1. *If a and b are unsigned numbers then:*

Case 1. $(+a)+(+b)=+(a+b)$;

Case 2. $(+a)+(-b)=(-b)+(+a)=+(a-b)$ if $a>b$;

Case 3. $(+a)+(-b)=(-b)+(+a)=-(b-a)$ if $a<b$;

Case 4. $(+a)+(-b)=(-b)+(+a)=0$ if $a=b$;

Case 5. $(-a)+(-b)=-(a+b)$;

Case 6. $(+a)+0=0+(+a)=+a$,

$(-a)+0=0+(-a)=-a$,

$0+0=0$.

Exercise. Substitute particular numbers in each case of definition 1 and verify that it yields the expected results. Interpret each illustration on Fig. 22.

It is convenient to identify our positive (signed) numbers with our previous unsigned numbers. We shall do this from here on. Hence we shall write 3 and + 3 interchangeably.

Our definition of multiplication of directed numbers will be motivated by our desire to have directed numbers behave as much like our former (unsigned) numbers as possible. For example, we would like them to obey the commutative and distributive laws.

Definition 1, section 13, tells us that $3 \cdot 2$ means $2+2+2$. Therefore we would like $3 \cdot (-2)$ to mean $(-2)+(-2)+(-2)$ or -6; that is, $3 \cdot (-2)$ means what you get by counting out three -2's and adding them up. However, we cannot similarly let $(-2) \cdot 3$ mean what you get by counting out -2 threes and adding them up. But, since we want the commutative law to remain in force, we would like $(-2) \cdot 3$ to be the same as $3 \cdot (-2)$, that is, -6. Similarly, we want $3 \cdot 0$ to mean $0+0+0$ or 0; because we want to keep the commutative law, we want $0 \cdot 3$ to be 0 as well.

As for the mysterious case $(-3)(-2)=+6$ we may motivate it by referring to the distributive law. Consider the expression $(-3)[(-2)+2]$. We want the distributive law to hold so that

$$(-3)[(-2)+2]=(-3)(-2)+(-3)\cdot 2.$$

But the left member is $(-3) \cdot 0$ which we should like to be zero. Hence the right member $(-3)(-2)+(-3) \cdot 2 = 0$. The second term, as we have already decided, should be -6. Hence $(-3)(-2)+(-6)=0$. For this to be true, $(-3)(-2)$ should be 6. Notice that we have not *proved* that $(-3)(-2)=6$. We have only shown that if we want the distributive law and other former rules to remain valid we have to *define* $(-3)(-2)=6$.

Another consideration that suggests this definition is the following. We would certainly like $(-3)(-2)$ to be either $+6$ or -6. Now $(-6)=(-3) \cdot 2$ by previous agreement. If we took $(-3)(-2)= -6$ and *if* we wanted the axiom that "if equals are divided by equals, the results are equal" to hold, then $(-3)(-2)=(-3) \cdot 2$ would yield $(-2)=2$ (dividing both sides by -3). This is undesirable.

Since we find it convenient to have our rules operate without exception we make the following definition of multiplication of directed numbers.

DEFINITION 2. *If a and b are unsigned numbers,*

$$
\begin{aligned}
(+a)(+b) &= +(ab),\\
(-a)(-b) &= +(ab),\\
(+a)(-b) &= (-b)(+a) = -(ab),\\
(+a) \cdot 0 &= 0(+a) = 0,\\
(-a) \cdot 0 &= 0 \cdot (-a) = 0,\\
0 \cdot 0 &= 0.
\end{aligned}
$$

This is sometimes called the "rule of signs."

Note that $2 \cdot 3 = 6$ is 3 less than $3 \cdot 3 = 9$, and $1 \cdot 3 = 3$ is 3 less than $2 \cdot 3 = 6$. Hence we would like $0 \cdot 3$ to be 3 less than $1 \cdot 3 = 3$, or 0; and we would like $(-1)3$ to be 3 less than $0 \cdot 3 = 0$, or -3; and we would like $(-2)3$ to be 3 less than $(-1)3 = -3$, or -6. Similarly we note that $1(-3)= -3$ is 3 more than $2(-3)=-6$. Hence we would like $0(-3)$ to be 3 more than $1(-3)=-3$, or 0; and we would like $(-1)(-3)$ to be 3 more than $0(-3)=0$, or $+3$; and we would like $(-2)(-3)$ to be 3 more than $(-1)(-3)=+3$, or $+6$. These desires for uniform behavior are satisfied by our definition 2.

Subtraction and division of directed numbers will be defined

as the inverses of addition and multiplication, respectively, as before.

DEFINITION 3. *If a and b are any directed numbers, then $a - b$ shall mean a directed number x such that $b + x = a$, if such a number exists. If there is a positive number x such that $b + x = a$, we shall say $a > b$ or $b < a$.*

Intuitively $a > b$ means that a is found to the right of b on the line in Fig. 22.

For example $(-2)-(-3)$ means a number x such that $(-3)+x = -2$; thus $x = 1$, since $(-3)+1 = -2$; hence $(-2)-(-3) = 1$. Similarly $2-(-3)$ means a number x such that $(-3)+x = 2$; thus $x = 5$, since $(-3)+5 = 2$; hence $2-(-3) = 5$. The results obtained here from our definition 3 are exactly what would be obtained from the familiar rule, learned by rote in high school: to subtract one quantity from a second, change the sign of the first and add.

Note that the minus sign is being used in two different ways: to indicate negative numbers, and to indicate the operation of subtraction. It might be preferable to use $\llcorner 2$ instead of -2 for the negative number "minus two" and reserve the minus sign for the operation of subtraction. Then $2-(-3) = 5$ would be written $2 - \llcorner 3 = 5$. However, we shall adhere to the customary notation.

It is intuitively evident and it can be proved that the difference between any two directed numbers *always exists* and is unique.

DEFINITION 4. *If a and b are any directed numbers, $a \div b$ shall mean the unique directed number x such that $bx = a$, if such a number exists.*

It can be shown that, except for the case where $b = 0$, to be discussed in the next section, the quotient $a \div b$ of any two directed numbers exists and is unique.

Examples. $(-6)\div 2$ means a directed number x such that $2x = -6$; thus $x = -3$, since $2(-3) = -6$; hence $(-6)\div 2 = -3$. Similarly, $6 \div(-2)$ means a number x such that $(-2)x = 6$; thus $x = -3$, since $(-2)(-3) = 6$; hence $6 \div(-2) = -3$. Finally, $(-6)\div(-2)$ means a number x such that $(-2)x = -6$; thus $x = 3$, since $(-2)3 = -6$; hence $(-6)\div(-2) = 3$.

EXERCISES

Calculate, using the appropriate definitions:

1. $5+(-2)$. **2.** $5-(-2)$. **3.** $(-5)+2$. **4.** $(-5)+(-2)$.
5. $(-5)-(-2)$. **6.** $(-5)-2$. **7.** $(-5)\cdot 2$. **8.** $5\cdot(-2)$.
9. $(-5)\cdot(-2)$. **10.** $(-8)\div 2$. **11.** $(-8)\div(-2)$. **12.** $8\div(-2)$.

DEFINITION 5. *The **negative** of a directed number x is a directed number having the property that when it is added to x the sum is zero. The negative of a directed number x is denoted by $-x$. In other words, $-x$ stands for the number $0-x$.*

Example. The negative of $(+3)$ is (-3) since $(+3)+(-3)=0$. Similarly the negative of (-3) is $(+3)$ since $(-3)+(+3)=0$. Hence $-(+3)=-3$ and $-(-3)=+3$. That is, $+3$ and -3 are negatives of each other.

In general, x and $-x$ are negatives of each other; thus $-(-x)=x$.

That is, if a directed number x is interpreted as a marching order, starting from 0, then $-x$ shall signify what you must do to get back to 0 after obeying the order "x." Thus, the order "$-x$" undoes the order "x."

This is still a third meaning for the minus sign. It can be shown that no trouble arises from this triple use of the minus sign.

We could now prove that the associative, commutative, and distributive laws hold for directed numbers; we shall not do so here. All the rules for manipulating fractions carry over to directed numbers in an obvious way. For example, $\frac{ax}{bx}=\frac{a}{b}$ where a, b, x are any directed numbers (b and $x\neq 0$). The fraction line is used interchangeably with the $\div$ sign. In short, *our definitions enable you to manipulate directed numbers as you have been taught to do it in high school.* While we shall not establish all this in detail here, we prove the following theorems as illustrations; each of them justifies a manipulation long familiar to the student.

THEOREM 1. *If a and b are any directed numbers, then $a-b=a+(-b)$.*

Proof. Note that the minus sign on the left indicates the operation of subtraction, while the one on the right indicates the

negative of b (def. 5). By definition 3, we have only to show that b plus the right member equals a. But

$$
\begin{aligned}
b+[a+(-b)] &= [a+(-b)]+b && \text{(commutative law for addition)}\\
&= a+[(-b)+b] && \text{(associative law for addition)}\\
&= a+0 && \text{(definition 5)}\\
&= a && \text{(definition 1).}
\end{aligned}
$$

This completes the proof.

Theorem 1 says that subtraction of b is equivalent to addition of $-b$. To apply the commutative law for addition, say, to $a-b$ it is understood that we are to write the expression as a sum first; thus, $a-b = a+(-b)=(-b)+a$.

Theorem 2. *If a and b are any directed numbers, then* $a-(-b)=a+b$.

Proof. By theorem 1, a minus $(-b)$ is equal to a plus the negative of $(-b)$. But the negative of $(-b)$ is b. Hence, $a-(-b)=a+b$.

Theorem 3. *If a and b are any directed numbers, then* $a(-b) = -(ab)$.

Proof. By definitions 5 and 2,

$$a[b+(-b)]=a\cdot 0=0.$$

By the distributive law,

$$a[b+(-b)]=ab+a(-b).$$

Hence,

$$ab+a(-b)=0.$$

By definition 5, this means that $a(-b)$ is the negative of ab. This completes the proof.

Theorem 4. *If a and b are any directed numbers, then* $(-a)(-b)=ab$.

Proof. By theorem 3, $(-a)(-b)$ equals the negative of $(-a)b$. But $(-a)b = -(ab)$ by theorem 3 and the commutative law for multiplication. Hence the negative of $(-a)b$ equals the negative of $-(ab)$, that is, ab. Hence, $(-a)(-b)=ab$.

It can also be proved that $\dfrac{-a}{b}=\dfrac{a}{-b}=-\dfrac{a}{b}$ and $\dfrac{-a}{-b}=\dfrac{a}{b}$, where a and b are any signed numbers, $b\neq 0$. Compare exercises 13 and 27 below.

We shall use these rules for manipulating directed numbers in the remainder of the book without explicit mention.

Our rules for manipulating directed numbers fit in well with applications like the following.

Example 1. Consider the rectangles in Fig. 23. The area of the unshaded rectangle is $(a - c)(b - d)$, or $[a + (-c)][b + (-d)]$. Applying the distributive law to this expression we get $[a + (-c)]\,b + [a + (-c)]\,(-d) = ab + (-c)b + a(-d) + (-c)(-d)$. Now using our definitions for multiplying directed numbers, we get $ab - cb - ad + cd$ as the area of the unshaded rectangle. This is in agreement with the geometry of the situation for ab is the area of the entire large rectangle, cb is the area of the shaded horizontal rectangle, ad is the area of the shaded vertical rectangle, and cd is the area of the cross-shaded rectangle. Our result is in agreement with the geometric interpretation of each term, because in subtracting cb and ad from ab we have subtracted the cross-shaded area twice and therefore have to put it back once by adding the term cd. However, this *does not prove* that the definition $(-c)(-d) = cd$ *must* be adopted. It merely shows that *if* we want our multiplication of directed numbers to be applicable to the geometrical situation above and *if* we want the distributive law to hold for directed numbers, *then* we should make the *definition* $(-c)(-d) = cd$.

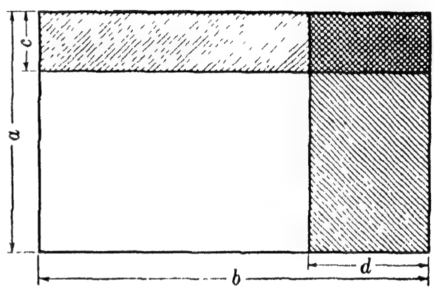


Fig. 23

Example 2. If the water level in a tank is rising at the steady rate of 2 inches per minute, and we want to know how much higher it will be in 3 minutes from now, the answer is $3 \cdot 2$ or 6 inches. The usual rule is $rt = a$ (rate times time equals amount). Suppose we interpret negative time as referring to the past and negative rate as meaning that the water level is sinking. We can say that if the water level is changing at the rate of 2 inches per minute, then 3 minutes ago it was $2(-3) = -6$ inches "higher" or 6 inches lower. If the water level is sinking at the rate of 2 inches per minute, we might say it is changing at the rate of -2

inches per minute. Then 3 minutes from now it will have changed $(-2)\cdot 3 = -6$ inches; that is, it will have sunk 6 inches. Similarly, 3 minutes ago (that is, -3 minutes from now), it would be $(-2)(-3) = +6$ inches higher. Our rules for directed numbers evidently fit this practical situation. But we have *not proved* here that our rules are correct. We have only shown that *if* we want our rules for negative numbers to fit this application and *if* we want the formula $rt = a$ to apply in all cases, *then* we should *define* $(-2)(-3) = 6$.

EXERCISES

13. Prove that (*a*) $\frac{-6}{3} = -2$; (*b*) $\frac{6}{-3} = -2$; (*c*) $-\frac{6}{3} = -2$; (*d*) $\frac{-6}{-3} = 2$.

14. Does $-a$ always represent a negative number? Illustrate.

Calculate and simplify each of the following:

15. $\frac{1}{2} + \frac{-3}{2}$. **16.** $\frac{-1}{2} + \frac{-3}{2}$. **17.** $\frac{-1}{2} + \frac{2}{3}$. **18.** $\frac{1}{2} - \frac{2}{3}$.

19. $-\frac{1}{2} - \frac{2}{3}$. **20.** $\left(-\frac{1}{2}\right)\left(\frac{3}{2}\right)$. **21.** $\left(\frac{-1}{2}\right)\left(\frac{-3}{2}\right)$.

22. $\left(-\frac{1}{2}\right) \div \frac{3}{2}$. **23.** $\frac{1}{2} \div \left(-\frac{3}{2}\right)$. **24.** $\left(-\frac{1}{2}\right) \div \left(\frac{-3}{2}\right)$.

25. By means of the generalized distributive law, simplify:
(*a*) $(-1)(a-b)$; (*b*) $(-1)(-a+b)$.

26. Prove, by means of definition 5, that:
(*a*) $-(a-b) = -a+b$; (*b*) $-(-a+b) = a-b$.

27. Prove that:
(*a*) $\frac{-a}{b} = \frac{a}{-b}$; (*b*) $\frac{-a}{b} = -\frac{a}{b}$; (*c*) $\frac{-a}{-b} = \frac{a}{b}$;
where a and b are any signed numbers, $b \neq 0$.

28. Show that $[3+(-2)]+(-4) = 3+[(-2)+(-4)]$.

29. Show that $(-3)[5+(-2)] = (-3)5+(-3)(-2)$.

22. The system of rational numbers. All the numbers introduced so far, namely zero, the positive and negative integers, and the positive and negative fractions, (and no others), are called **rational numbers.** The word "rational" does not mean reasonable. It comes from the word ratio; every rational number can be expressed as a quotient or ratio of two integers.

When we had natural numbers only, the sum and product of two natural numbers was always a natural number, but we could not say as much about quotients and differences. With fractions we could say that the sum, product, or quotient of any two fractions was always a fraction, but we could not say as much for differences. But the system of all rational numbers is closed under addition, subtraction, multiplication and division save for one exceptional case. That is, it can now be proved that the sum, difference, product and quotient of any two rational numbers exists within the system of rational numbers, with the single exception that *division by zero must be excluded.* Why we must make this exception will be seen at once from the definition of division:

The quotient $a \div b$ *is the unique number* x *such that* $bx = a$, *if such a number exists.*

Let us attempt to apply this definition with $b = 0$. Now either $a \neq 0$ or $a = 0$.

Case I. Let $a \neq 0$. Then $a \div 0$ means a number x such that $0 \cdot x = a$. But $0 \cdot x = 0$ no matter what number x is and $a \neq 0$. Hence the definition cannot be satisfied. For example, $3/0$ means a number x such that $0 \cdot x = 3$, provided such a number x exists. But clearly no such number can exist since $0 \cdot x = 0$ and not 3 no matter what x is.

Case II. If $a = 0$, $a \div 0$ or $0 \div 0$ means a number x such that $0 \cdot x = 0$; but this equation is satisfied by any number x whatever and hence is quite useless.

Therefore division by zero is excluded in all cases.

This is analogous to the following situation. Suppose the only requirement for admission to the college were that one must be more than 30 ft. tall. Then no one could satisfy the requirement. This is analogous to Case I where no number can satisfy the requirement of the definition. Case II is analogous to the situation we would find if the only entrance requirement were that one must be less than 30 ft. tall. Then the requirement excludes no candidate. Neither arrangement is a usable requirement.

On the other hand, $0 \div b$ does have a definite value if $b \neq 0$. By definition $0 \div b$ means a number x such that $bx = 0$. There is such a number x, namely zero. *Thus* $0 \div b = 0$ *if* $b \neq 0$. For

example, $0 \div 3$ means a number x such that $3x = 0$; hence $x = 0$. That is $0 \div 3 = 0$.

Note that $0/1 = 0$ while $1/0$ is a meaningless symbol, since there exists no number satisfying the definition of $1/0$. To say that a symbol has no meaning is not the same as to say it has the meaning "zero." A student who is not registered for this course receives no mark in it; a student must register for the course before he can aspire to the mark of zero.

Amusing results can be obtained if division by zero is overlooked. (We assume for the moment that the student recalls some high school algebra.) For example, let Alice be a years old and Betty be b years old. Suppose the two girls are of the same age. Then $b = a$. Multiplying both sides of this equation by a, we obtain $ab = a^2$. Subtracting b^2 from both sides of the equation, we obtain $ab - b^2 = a^2 - b^2$. Factoring, we have $b(a - b) = (a + b)(a - b)$. Dividing both sides by $(a - b)$ we have $b = a + b$, or $b = a + a$, or $b = 2a$. Thus Betty discovers to her dismay that she is twice as old as Alice. The trouble is, of course, that we have divided by $a - b$ which is zero since $a = b$ by hypothesis, and division by zero has been excluded. If division by zero were permissible we could obtain a somewhat oversimplified system of arithmetic which might well appeal to some students although it would not be very practical. For, from the result $b = 2a$ just obtained above, we get $a = 2a$ and hence, dividing by a, we obtain $1 = 2$. Then $3 = 2 + 1 = 1 + 1 = 2 = 1$, and $4 = 3 + 1 = 1 + 1 = 2 = 1$, and so on. Hence all numbers would be equal to each other and there would be no such thing as a wrong answer.

As remarked above, the quotient of any two rational numbers exists except when the divisor is zero. In particular $1/a$ exists if $a \neq 0$.

Definition 1. *If $a \neq 0$, the number $1/a$ is called the* ***reciprocal*** *of a.*

Note that division by $a(\neq 0)$ is equivalent to multiplication by $1/a$.

The number zero has another important property, which will be useful later, as follows.

Theorem 1. *If $ab = 0$ and $a \neq 0$ then b must be 0.*

Proof. Since $a \neq 0$, there is a rational number $1/a$. By hypothesis, $ab = 0$. Multiplying both sides by $1/a$, we have by E_6, that

$$\frac{1}{a} \cdot ab = \frac{1}{a} \cdot 0. \tag{1}$$

By our now familiar processes of algebra, the left member becomes $\dfrac{ab}{a}$ or b and the right member becomes 0. Hence (1) becomes $b = 0$, which is what we had to prove.

This theorem may be stated in other ways, as follows: *the product of two non-zero numbers cannot be zero; if the product of two numbers is zero, at least one of the factors is itself zero; the product of two numbers can be zero only when one or the other (or both) of the factors is zero.*

We may characterize rational numbers by means of the following theorem, which we shall not prove here.

A number is **rational** *if and only if it can be expressed as the quotient a/b of two integers a and b, with $b \neq 0$.*

Thus $\dfrac{-7}{2}$ and $\dfrac{3}{-21}$ are rational numbers.

We can classify rational numbers as follows:

Rational Numbers (or positive and negative fractions and zero)
- *Integers* (rational numbers which can be expressed with denominators = 1)
  - *Negative integers*
  - *Zero*
  - *Positive integers* or *Natural Numbers.*

EXERCISES

1. If a and b are any two rational numbers, which of the following expressions must also represent a rational number:

(*a*) $a + b$; (*b*) $a - b$; (*c*) $a \cdot b$; (*d*) $a \div b$?

2. If a and b are any two integers, which of the expressions in exercise 1 must also represent an integer?

23. Powers and roots. The following definition is no more than an abbreviation.

DEFINITION 1. *If n is a natural number, x^n means $x \cdot x \cdots x$ with n factors.*

For example, $x^3 = xxx$. The **exponent** n is merely the number of factors. We say that x^n is the ***n*th power** of x. x^2 is called the **square** of x, and x^3 is the **cube** of x.

DEFINITION 2. *If there exists a number x such that $x^n = a$, n being a natural number, then x is called an **nth root** of a. In particular if $x^2 = a$, x is called a **square root** of a, and if $x^3 = a$, x is called a **cube root** of a.*

For example, 3 is a square root of 9 since $3^2 = 9$. Similarly (-3) is a square root of 9 since $(-3)^2 = 9$. Both 2 and -2 are fourth roots of 16 since $2^4 = 16$ and $(-2)^4 = 16$. A cube root of 8 is 2.

If a is a positive number, we shall use the radical sign $\sqrt[n]{a}$ to denote the positive nth root of a, provided such a positive nth root exists. The number n written above the radical sign is called the **index** of the root. In the case of square roots it is customary to omit the index 2. Thus $\sqrt{9}$ stands for 3, but not for -3. If we wish to indicate -3 we shall write $-\sqrt{9}$. This agreement is made in order to avoid the confusion that would arise if we allowed the symbol $\sqrt{9}$ to stand for either 3 or -3 ambiguously. Similarly, although 2 and -2 are both fourth roots of 16, the symbol $\sqrt[4]{16}$ will mean 2 alone. We shall be almost always concerned with nth roots of positive numbers, only; these are unique whenever they exist.

We make the linguistic agreement that *powers shall take precedence over multiplication and division, except where parentheses indicate otherwise.* For example, $5 \cdot 2^3 = 5 \cdot 8 = 40$ while $(5 \cdot 2)^3 = 10^3 = 1000$. *The radical sign, however, acts as a parenthesis; all indicated operations under the radical sign are to be done before extracting the root.* For example, $\sqrt{9 + 16}$ means $\sqrt{(9 + 16)}$ $= \sqrt{25} = 5$.

It may be conjectured that our radical sign (from the Latin word "radix" meaning "root") comes from the letter r and the "vinculum" $\overline{}$ (a Latin word meaning "bond") which was often used as we use parentheses. For example, older books might write $a \cdot \overline{b + c} = ab + ac$ for our distributive law. Hence $\sqrt{x + y}$ means $\sqrt{(x + y)}$.

THEOREM 1. *If n is any positive integer then $(xy)^n = x^n y^n$.*

Proof. $(xy)^n = (xy)(xy) \cdots (xy)$ where there are n parentheses. By the generalized associative law for multiplication we have

$$\begin{aligned}(xy)^n &= xyxy \cdots xy \\ &= xx \cdots xyy \cdots y && \text{(commutative law for multiplication)} \\ &= x^n y^n && \text{(by definition 1).}\end{aligned}$$

For example, $(xy)^3 = (xy)(xy)(xy) = xxxyyy = x^3y^3$.

THEOREM 2. *If n is any positive integer, then $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$, provided that all these nth roots exist, a and b being positive.*

Proof. Let $x = \sqrt[n]{a}$; then $x^n = a$ by definition 2. Let $y = \sqrt[n]{b}$; then $y^n = b$. By theorem 1, $(xy)^n = x^n y^n = ab$. Hence $xy = \sqrt[n]{ab}$ by definition 2. By substitution we have $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$.

For example, $\sqrt{4} \cdot \sqrt{9} = \sqrt{4 \cdot 9} = \sqrt{36}$.

EXERCISES

1. Find (*a*) $\sqrt[3]{27}$; (*b*) $\sqrt[4]{16}$; (*c*) $\sqrt{49}$; (*d*) $\sqrt[5]{32}$; (*e*) $\sqrt[6]{64}$; (*f*) $\sqrt[4]{81}$.

2. Is it true that $\sqrt{a^2 + b^2} = a + b$? Illustrate with particular numbers replacing a and b.

3. Is it true that $\sqrt{a^2 - b^2} = a - b$? Illustrate with particular numbers replacing a and b.

4. Of what is $a + b$ a square root? (We assume some previous knowledge of algebra here.)

5. Of what is $a - b$ a square root? (We assume some previous knowledge of algebra here.)

6. Is it true that $\sqrt{x + y} = \sqrt{x} + \sqrt{y}$? Illustrate with particular numbers replacing x and y.

7. Is it true that $\sqrt{x - y} = \sqrt{x} - \sqrt{y}$? Illustrate with particular numbers replacing x and y.

Simplify each of the following:

8. $(\sqrt{5})^2$.

9. $\sqrt{\dfrac{1 - \dfrac{1}{2}}{2}}$.

10. $\sqrt{36 + 64}$.

11. $\sqrt{1 - \left(\dfrac{\sqrt{3}}{2}\right)^2}$.

12. $(\sqrt[3]{a})^3$.

13. $(\sqrt[5]{x^2})^5$.

14. $(\sqrt[p]{x})^p$. **15.** $4 \cdot 3^2$. **16.** $(4 \cdot 3)^2$.
17. $5 + 2 \cdot 3^2$. **18.** $(5 + 2) \cdot 3^2$. **19.** $5 + (2 \cdot 3)^2$
20. $(5 + 2 \cdot 3)^2$. **21.** $([5 + 2]3)^2$. **22.** $\sqrt{36\,x^2}$.
23. $\sqrt{3\,x} \cdot \sqrt{12\,x}$. **24.** $\sqrt{2\,xy}\sqrt{8\,xy}$. **25.** $(\sqrt{3\,x})^2$.

24. The square root of two. We shall prove the following surprising theorem.

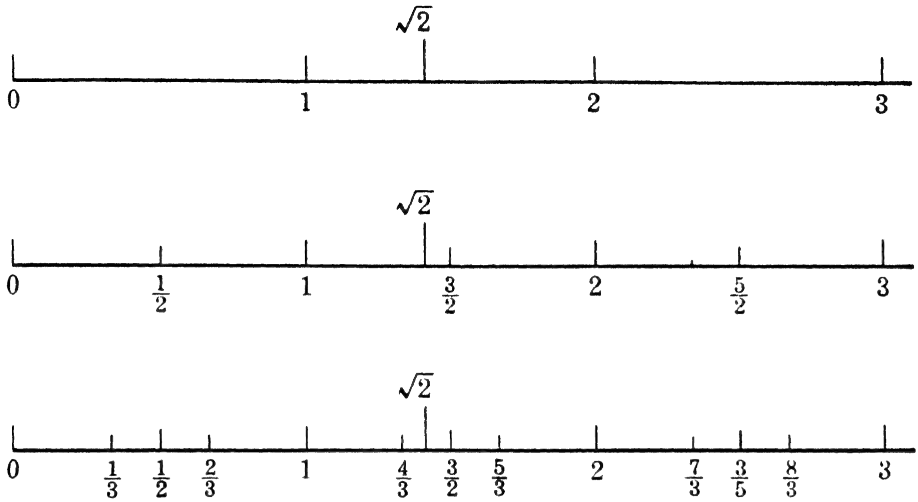
THEOREM 1. *No rational number can be a square root of 2.*

Before proving this theorem, let us see why it is surprising. A rational number is one which can be expressed as a quotient a/b of two integers ($b \neq 0$). If we arrange the positive rational numbers according to the following scheme

$$\begin{array}{cccccc}
\frac{1}{1} & \frac{2}{1} & \frac{3}{1} & \frac{4}{1} & \frac{5}{1} & \cdots \\
\frac{1}{2} & \frac{2}{2} & \frac{3}{2} & \frac{4}{2} & \frac{5}{2} & \cdots \\
\frac{1}{3} & \frac{2}{3} & \frac{3}{3} & \frac{4}{3} & \frac{5}{3} & \cdots \\
\frac{1}{4} & \frac{2}{4} & \frac{3}{4} & \frac{4}{4} & \frac{5}{4} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots &
\end{array}$$

we see at once that every positive rational number is represented somewhere in this endless array. For example, the number 273/565 is to be found in the 273rd column and 565th row. Let us imagine the points corresponding to these numbers marked off on a line one row at a time, the number attached to each point representing its distance from the zero-point or origin. At each stage (Fig. 24) the marked points are more and more thickly strewn ("densely distributed" is a technical term for this) on the line. Although it cannot * be accomplished in practice, since the number of rational numbers is unlimited, we can imagine all the rational numbers marked off on the line. You might think that they fill up the entire line. But if our theorem is correct, this is not so; despite the density of the marked points on the line there are points which are not marked. One such point can be con-

* However any particular fraction can be marked off.



The first three stages

FIG. 24

structed by taking an isosceles right triangle whose side is 1 unit in length. Its hypotenuse then must be $\sqrt{1^2 + 1^2} = \sqrt{2}$ by the Pythagorean theorem. The length of this hypotenuse can then be laid off on the line from the origin and the other end cannot be a marked point, if our theorem is correct. For all marked points have rational numbers corresponding to them, while our theorem asserts that $\sqrt{2}$, whatever it may be, is not a rational number. If you recall that we have assumed that *all* rational numbers, including all fractions with denominators equal to a million, a billion, a trillion, etc., have been marked off, it may seem remarkable that there is still a "gap," that is, an unmarked point, left. The density of the rational numbers may be seen from the fact that between any two rational numbers a and b no matter how close together, there must be another rational number; in fact, $\dfrac{a+b}{2}$ is such a rational number. It is not hard to see intuitively that $\dfrac{a+b}{2}$ is exactly midway between a and b.

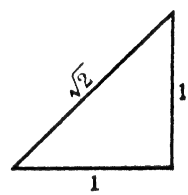


FIG. 25

Let us now look at theorem 1 from a somewhat different point of view. Consider two line-segments. A length is said to be a **common measure** of these two line-segments if it goes into *each* of them *exactly* a *whole* number of times. For example, if the two

given line-segments are 1 yd. and $1\frac{1}{4}$ ft. respectively, then 1 inch is a common measure for them, since 1 inch goes into the first 36 times and into the second 15 times. Similarly, a 1 ft. line-segment and a $5\frac{3}{4}$ in. line-segment have $\frac{1}{4}$ in. as a common measure, since $\frac{1}{4}$ in. goes into the first 48 times and into the second 23 times. If two line-segments have a common measure they are called **commensurable.** It would seem intuitively that any two line-segments would be commensurable if only we try a small enough length as common measure. But if theorem 1 is correct, this is not the case. The leg and hypotenuse of the right triangle in Fig. 25 can have no common measure. For, suppose they could have a common measure. Then it would go into the unit leg q times and into the hypotenuse p times, where q and p are both whole numbers. But then the common measure would be represented by the number $1/q$ and the hypotenuse by the *rational* number p/q. But, according to theorem 1, this is impossible since $\sqrt{2}$ is not a rational number. Two line-segments which have no common measure are called **incommensurable.** The existence of incommensurable lengths was probably known to Pythagoras (about the 6th century B.C.) and is said to have disturbed him so much that he tried to suppress the information lest it discredit mathematicians in the eyes of the general public.

Still another way to understand what our theorem asserts is to imagine all the rational numbers enumerated in some order. If we tried to enumerate them one row at a time we would never get to the second row since we could never finish the first row. Hence we enumerate them in the diagonal order indicated in Fig. 26 by the arrows, starting with $1/1$. That is, we write them in the following order:

$$(A)\quad \frac{1}{1},\ \frac{1}{2},\ \frac{2}{1},\ \frac{3}{1},\ \frac{2}{2},\ \frac{1}{3},\ \frac{1}{4},\ \frac{2}{3},\ \frac{3}{2},\ \frac{4}{1},\ \cdots$$

In this way we must reach any given positive rational number if we proceed along the sequence (A) far enough. Our theorem asserts that none of them will yield exactly 2 when squared. This obviously cannot be proved by trying them in succession, be-

cause life is short and there is no end to the sequence (A) of positive rational numbers. If there were only a limited number of rational numbers we could prove the theorem by testing each number. But the number of rational numbers is unlimited. If we squared the first thousand (or first billion) positive rational numbers in the sequence (A) without getting 2 as the result, we could not be sure that we might not get 2 at some later trial. It might be that we simply hadn't found the right one yet.

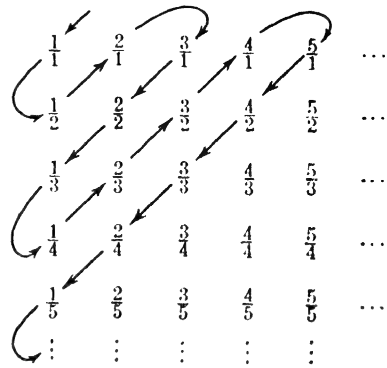


FIG. 26

How then *can* we prove that no rational number can yield exactly 2 when squared? We shall need a few preliminaries.

DEFINITION 1. *An integer is called* **even** *if and only if it can be expressed as* $2x$ *where* x *is some integer.*

In other words, an integer is even if it is divisible by 2, or has 2 as a factor. For example, 10 is even since $10 = 2 \cdot 5$.

Remark. Notice the necessary restriction that x be an integer. The number 3 can be expressed as $2 \cdot (3/2)$ but we would not call 3 even.

An integer which is not even is called **odd.** It can be proved that *an integer is odd if and only if it can be expressed as* $2x + 1$ *where* x *is an integer.* In other words every odd integer is exactly *one* more than some even integer. We shall not prove this here. For example, $7 = 2 \cdot 3 + 1$, and $11 = 2 \cdot 5 + 1$ are odd. Or, an integer is odd if you get a remainder of one when you divide it by 2.

* THEOREM A. *If* a *is even,* a^2 *is even.*

Proof. Since a is even we can write $a = 2x$ where x is an integer, by definition 1. Thus $a^2 = (2x)^2 = 4x^2 = 2(2x^2)$. Now since x is an integer so is x^2, because the product of two integers is an integer, and hence so is $2x^2$ for the same reason. Hence a^2 has been expressed as 2 times the *integer* $2x^2$. Thus a^2 is even, by definition 1.

* Theorem A is not needed, but is included to emphasize the distinction between it and theorems B and C.

THEOREM B. *If a is odd, a^2 is odd.*

Proof. Since a is odd we can write $a = 2x + 1$ where x is an integer. Thus $a^2 = (2x+1)^2 = 4x^2 + 4x + 1 = 2(2x^2 + 2x) + 1$. (We assumed in this step that the student remembers how to multiply $2x + 1$ by itself. This will be taken up again in Chapter V. Compare exercise 7, section 14.) Now since x is an integer and since the sum and product of two integers is again an integer, $2x^2 + 2x$ is an integer. Hence a^2 has been expressed as twice the *integer* $(2x^2 + 2x)$ plus 1 and is therefore odd.

THEOREM C. *If n is an integer and n^2 is even, then n is even.*

Proof. Since n is an integer it is either even or odd. If n were odd, then n^2 would have to be odd by Theorem B. Since n^2 is even, by hypothesis, this cannot be so. Hence the only possibility is that n is even, which is what we had to prove.

We are now ready for the proof of Theorem 1.

Proof of Theorem 1. Either there is a rational number $=\sqrt{2}$ or there is not. Suppose there were such a rational number; it could be reduced to lowest terms, and hence expressed as p/q where p and q are integers having no factor in common except ± 1 $(q \neq 0)$. Then,

$$p/q = \sqrt{2}.$$

If this were so, then

$$p^2/q^2 = 2. \qquad \text{(Reason?)}$$

Hence

$$p^2 = 2q^2. \qquad \text{(Reason?)} \tag{1}$$

Now q is an integer; therefore q^2 is an integer. Hence (1) says that p^2 is an even integer since it is twice the integer q^2. By Theorem C it follows that p is an even integer. Since p and q have no factor in common q *must be odd,* for if q were also even, p and q would have a common factor 2. Since p is even, we can write $p = 2x$ where x is an integer. Substituting this in (1) we get

$$(2x)^2 = 2q^2$$

or,

$$4x^2 = 2q^2$$

or,

$$q^2 = 2x^2. \qquad \text{(Reason?)}$$

But this says that q^2 is even since it is twice the integer x^2. By Theorem C it follows that *q is even.* But the italicized statements contradict each other since q cannot be *both* even *and* odd; hence the supposition that $p/q = \sqrt{2}$ is absurd because it leads to a false conclusion. The only remaining possibility is that no rational number can be $= \sqrt{2}$, which is what we had to prove.

Remark. In this section we used the so-called *indirect proof* or *reductio ad absurdum* with which you first became acquainted in your high school geometry. That is, we show that since supposition A leads logically to a false (or, self-contradictory) conclusion B then A must be false. This is merely the logical principle that if A implies B and B is false then A is false (see section 4, exercise 14).

EXERCISES

1. Prove that if a is even, then a^3 is even. (Hint: imitate as far as possible, the proof of Theorem A.)

2. (*a*) Prove that if a is odd, then a^3 is odd. (Hint: imitate as far as possible the proof of Theorem B.)

(*b*) Prove that if n is an integer and n^3 is even, then n is even. (Hint: imitate as far as possible the proof of Theorem C. Use the theorem of exercise 2*a*.)

(*c*) Prove that no rational number is equal to $\sqrt[3]{2}$. (Hint: using exercise 2*b*, imitate as far as possible the proof of Theorem 1.)

3. Prove that if a is divisible by 3 then a^2 is divisible by 3. (Hint: by definition, an integer is divisible by 3 if and only if it can be expressed as $3x$ where x is some integer.)

4. (*a*) Prove that if a is not divisible by 3, then a^2 is not divisible by 3. (Assume that if an integer is not divisible by 3 it can be expressed as either $3x + 1$ or $3x + 2$ where x is an integer; that is, if you divide any integer by 3, you get a remainder of 0, 1, or 2. For example, $15 = 3 \cdot 5$, $16 = 3 \cdot 5 + 1$, $17 = 3 \cdot 5 + 2$, $18 = 3 \cdot 6$, $\cdots$. Imitate the proof of Theorem B as far as possible. There will be two cases.)

(*b*) Prove that if n is an integer and n^2 is divisible by 3 then n is divisible by 3. (Hint: Imitate as far as possible the proof of Theorem C. Use exercise 4*a*.)

(*c*) Prove that no rational number is equal to $\sqrt{3}$. (Hint: using exercise 4*b*, imitate as far as possible the proof of Theorem 1.)

5. Prove that no rational number is equal to $\sqrt[4]{2}$.

6. Prove that no rational number is equal to $\sqrt{5}$.

7. Prove that no rational number is equal to $\sqrt[3]{3}$.

8. Prove that the sum of two even numbers is even.

9. Prove that the product of two odd numbers is odd.

10. Prove that the product of any natural number and an even number is even.

11. Prove that no rational number is equal to $\sqrt{6}$. (Hint: use exercise 9.)

25. Irrational numbers. Decimal notation. The fact that no rational number is exactly equal to $\sqrt{2}$ means that our system of rational numbers contains no number which can represent the length of the hypotenuse of an isosceles right triangle whose leg is of unit length. This is regarded as an inadequacy in the rational number system. We are therefore led to invent *irrational numbers.* The word "irrational" must not be taken to mean "unreasonable"; it means not expressible as a ratio of two integers. We shall give no accurate definition of irrational number here because of the technical difficulties involved. In fact, as we remarked above, the strange story of $\sqrt{2}$ was probably known to Pythagoras (about the 6th century B.C.) but the difficulties involved were not straightened out with complete logical rigor until about 1870 when G. Cantor and R. Dedekind, two German mathematicians, did it independently of each other. However, we can point out some of the essential characteristics of irrational numbers.

Notice that the usual decimal notation for numbers, which you learned as children, is based on powers of 10. For example, 2347.568 *means*

$$2 \cdot 10^3 + 3 \cdot 10^2 + 4 \cdot 10 + 7 + \frac{5}{10} + \frac{6}{10^2} + \frac{8}{10^3}.$$

Recall the peculiar process you once memorized for extracting the square root of a number as a decimal. (Your knowledge of this weird process is a splendid example of sheer memorization without the slightest understanding. In fact, it will be unnecessary to recall the process as far as this book is concerned for, as we shall see soon, the same results can be obtained in a more straightforward, although slower, way.) Applying this process to the task of extracting the square root of 2, we get 1.4 as a first result, but there is a remainder. If we go further we get 1.41, then 1.414, then 1.4142, and so on. At each stage the process produces a remainder and hence fails to give an exact square root of 2. In fact,

$$(1.4)^2 = 1.96 < 2 < (1.5)^2 = 2.25,$$
$$(1.41)^2 = 1.9881 < 2 < (1.42)^2 = 2.0164,$$
$$(1.414)^2 = 1.999396 < 2 < (1.415)^2 = 2.002225,$$

and so on. Thus the decimal expression for $\sqrt{2}$ does not seem to stop. We cannot prove that it doesn't stop by continuing the process until we are exhausted for if it hadn't stopped by the 1000th place, we could still not be sure that it wouldn't stop at the 2000th. But we *can* in fact prove that the process *never* stops. For if it did stop we would have a terminated decimal expression for $\sqrt{2}$. Now ***if a decimal terminates then it represents a rational number;*** that is, the quotient of two integers. For example 2.315 means $\frac{2315}{1000}$. But $\sqrt{2}$ is not a rational number, by section 24. Hence its decimal expression cannot stop. However, each of the successive stages of the process gives a rational number whose square becomes closer and closer to 2. That is, $\sqrt{2}$ *is approximated by rational numbers and can be so approximated as closely as we wish* (that is, to the nearest thousandth, or millionth, etc.) by carrying the process far enough. *This is one essential characteristic of irrational numbers.*

The converse of the theorem in heavy type above would say that if a number is rational, then its decimal expression terminates. *This is not so,* as can be seen from the rational numbers

$$1/3 = .3333333\cdots \quad \text{and}$$
$$1/7 = .\underline{142857}\,\underline{142857}\,\underline{142857}\cdots$$

whose decimal expressions do not terminate, as you can see from the process of obtaining them by division. Notice that both of these decimals "repeat in blocks"; in technical language they are *periodic.* In the case of 1/7, notice that when you divide 1 by 7, the only remainders which can occur in the process of long division are the numbers 0, 1, 2, 3, 4, 5, 6. If 0 occurred, the process would terminate. If 0 does not occur, the only possible remainders are 1, 2, 3, 4, 5, 6. After enough steps (seven or fewer) are made in the process of long division, we must therefore get one of these remainders back again for the second time, whereupon the whole process repeats itself exactly, thus producing a periodic decimal.

Exercise. Perform the long division of 1 by 7 long enough to observe the described phenomenon.

In the same way it can be proved that if the decimal expression for any rational number p/q does not terminate, then it is neces-

sarily periodic. It can also be proved that every periodic decimal represents a rational number. It follows that the decimal expression for $\sqrt{2}$ can neither terminate nor be periodic. Every irrational number has a decimal expression which neither terminates nor is periodic.

We shall assume that the irrational numbers fill up all the "gaps" left in the line (Fig. 24), so that every point on the line has a number attached to it, rational or irrational. These numbers are called **real.** In particular, we now have a number to represent each length, such as, for example, the hypotenuse of Fig. 25. Every real number can be expressed as a decimal, either terminating or not, and conversely. The processes of addition, subtraction, multiplication and division can be defined for real numbers so that the sum, difference, product, and quotient of any two real numbers is again a real number, except that division by 0 remains excluded. The new numbers can be shown to satisfy the familiar laws, such as the associative, commutative, and distributive laws, and we are therefore content to call them numbers. We shall be unable to prove these statements here because we have not given a clear cut definition of real number. The system of real numbers is subdivided into irrational and rational numbers.

You might think that irrational numbers are rare because we have pointed out only a few. But it can be shown that the square root of any positive integer which is not the square of an integer is irrational. In general, the nth root of any positive integer which is not the nth power of an integer is irrational. There are many more irrational numbers which cannot be expressed as roots at all such as $\pi = 3.14159265 \cdots$, a number you met in your geometry course. It is not hard to prove the following exercises.

Exercise 1. Prove that if you add any rational number r to an irrational one, s, the sum t must be irrational. (Hint: the sum is either rational or not. Suppose it is rational and show that this leads to a contradiction.)

Exercise 2. Prove that if you multiply any rational number $r(\neq 0)$ by an irrational one, s, the product t is irrational.

Hence there are at least as many irrational numbers as rational numbers. It can be proved that, between every pair of rational numbers no matter how close together, there is an irrational one,

and between every pair of irrational numbers there is a rational one.

Whether a real number is rational or not does not matter to the practical engineer since measurements are never more than approximate and an irrational number can be approximated as closely as we please by rational numbers. But the idea of irrational numbers is essential to any careful theory of geometric magnitudes and to the logical derivation of the theorems of calculus, upon which much practical work is based. *All real numbers are either positive, negative, or zero, according to where they lie on the line* (Fig. 22); that is, a real number is positive or negative according as it lies to the right or the left of the origin or zero-point.

Note that it is not necessary to remember the complicated process for extracting square roots. One can proceed as follows. By

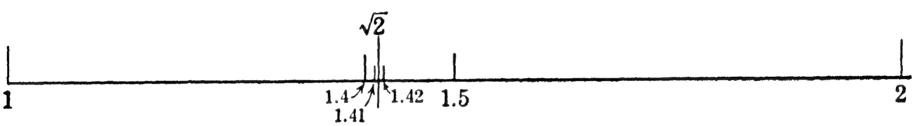


FIG. 27

multiplication we find that $1^2 = 1; 2^2 = 4$; and hence $1^2 < 2 < 2^2$, or,* $1 < \sqrt{2} < 2$. Then by trying successive tenths, $(1.1)^2$, $(1.2)^2$, and so on, we find, as above, that $(1.4)^2 < 2 < (1.5)^2$ or $1.4 < \sqrt{2} < 1.5$. Then by trying successive hundredths we find that $(1.41)^2 < 2 < (1.42)^2$ or $1.41 < \sqrt{2} < 1.42$; and so on. This simple straightforward method of successive approximations can be applied equally well to cube roots, fourth roots, etc. It resembles the process of running down a base-runner between first and second base in a baseball game, except that while we continue to pinch the irrational number between narrower and narrower limits, we never tag it.

Remark. The word "fraction" is often used for any indicated quotient as $\sqrt{2}/3$ which is not a fraction in the sense of section 16. To avoid confusion we might now refer to the "fractions" of section 16 as "positive rational fractions" or simply as "positive rational numbers."

* This depends on a theorem we shall not prove here, namely: if a and b are non-negative real numbers and $a < b$, then $\sqrt{a} < \sqrt{b}$.

EXERCISES

(*a*) *Approximate as far as three decimal places; that is, locate between successive thousandths:*

(*b*) *Write the answer correct to the nearest hundredth:*

1. $\sqrt{3}$. **2.** $\sqrt{5}$. **3.** $\sqrt[3]{2}$. **4.** $\sqrt[3]{3}$. **5.** $\sqrt[3]{5}$.

Show that each of the following numbers is irrational. (*Hint: assume it to be rational and show that this leads to a contradiction.*)

6. $1 + \sqrt{2}$. **7.** $3\sqrt{2}$. **8.** $\dfrac{\sqrt{2}}{3}$. **9.** $\dfrac{1 + \sqrt{2}}{4}$. **10.** $\dfrac{1 - \sqrt{2}}{4}$.

26. Complex numbers. Our original motive for the introduction of irrational numbers was our desire to have a number to represent each length, or to have a number to attach to every point on a line (Fig. 24). This implied in particular, that we wanted some number to be the square root of 2. While our number system consisted of only rational numbers, it contained no number exactly equal to $\sqrt{2}$, so we enlarged the number system again to include irrational numbers.

Suppose we now consider any negative number, say -4. A square root of -4 would be a number x such that $x^2 = -4$. We shall now prove that such a number x cannot be found even in our larger system of real numbers. For any real number x is either positive, zero, or negative. If x is positive, x^2 is also positive and cannot be -4. If $x = 0$, then $x^2 = 0$ and not -4. Finally, if x is negative, x^2 is positive and not -4. We have exhausted all the possibilities for a real number. This argument clearly applies to any negative number. Hence negative numbers have no square roots among the real numbers. In order to remedy this situation we take the symbol $\sqrt{-4}$ and call it a pure imaginary number. *Any symbol of the form* $b\sqrt{n}$ *where* b *stands for a real number* ($\neq 0$) *and* n *for any negative number, will be called a* ***pure imaginary*** *number.* For example $4\sqrt{-3}$ is a pure imaginary number. If we add a real number a to a pure imaginary number $b\sqrt{n}$ we get a peculiar hybrid beast $a + b\sqrt{n}$ (where a and b are real and n is negative) which is called a *complex* number. For example, $3 + 5\sqrt{-4}$ is a complex number.

The **complex numbers** $a + b\sqrt{n}$ (a and b being real and n being negative) *include* all the previous kinds of numbers as special cases. If $b \neq 0$ the complex number $a + b\sqrt{n}$ is called **imaginary.** For example, $3 + 5\sqrt{-4}$ is imaginary. If $a = 0$ and

$b \neq 0$ we have a *pure imaginary* number as $5\sqrt{-4}$. If $b = 0$ we have a *real* number. Thus 3 can be written as $3 + 0\sqrt{-1}$, and $3\sqrt{-2}$ as $0 + 3\sqrt{-2}$. We may summarize the entire number system as it now stands:

Complex Numbers $(a + b\sqrt{n})$
- *Imaginary* $(b \neq 0)$
  - *Pure Imaginary* $(a = 0,\ b \neq 0)$
- *Real* $(b = 0)$
  - *Irrational*
  - *Rational*
    - *Integers*
      - *Negative integers*
      - *Zero*
      - *Positive integers or Natural Numbers.*

Do not think that the word "imaginary" means that these numbers are mystical or fictitious or "unreal" in the everyday sense of the word, or that "complex" means complicated. It is true that pure imaginary and complex numbers did meet with opposition on some such grounds in the 16th and 17th centuries but so did negative numbers and with as little justice. Imaginary numbers have very "real" applications in many branches of physical science, into which we cannot go here, and may be given concrete interpretations. There is no room for the imaginary numbers on the line of Fig. 24, for the real numbers already account for every point on the line. To represent imaginary numbers graphically we use a whole plane. We shall not go into this here, since we shall not use imaginary numbers very much. As far as we are concerned, we are content to call them numbers for the usual reason that the familiar operations can be defined for them and that they can be shown to satisfy the familiar laws like the associative, commutative, and distributive laws. It is impossible to introduce a usable "less than" relation among all the complex numbers. Thus $<$ and $>$, and hence "positive" and "negative" are used only in connection with real numbers.

EXERCISES

1. List all the adjectives "complex, imaginary, pure imaginary, real, irrational, rational, integral, positive, negative, zero, natural" which apply to each number:

(*a*) $2 + 3\sqrt{-4}$. (*b*) $3\sqrt{4}$. (*c*) $3\sqrt{-4}$. (*d*) $3\sqrt{5}$.
(*e*) $\dfrac{\sqrt{2}}{3}$. (*f*) $-2/3$. (*g*) $(1 + \sqrt{3})/2$. (*h*) $(1 + \sqrt{9})/2$.

2. What was our original motive in introducing
(*a*) fractions; (*b*) negative numbers; (*c*) irrational numbers; (*d*) pure imaginary numbers; (*e*) complex numbers?

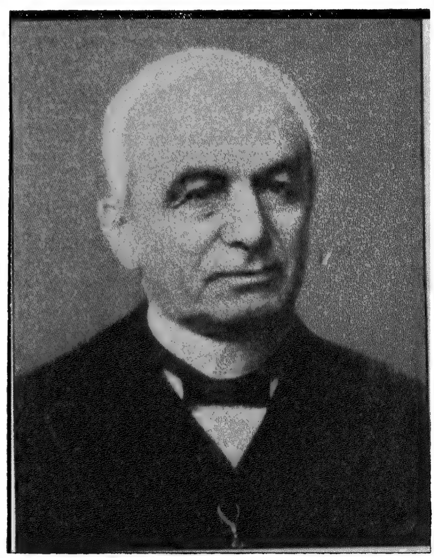
3. State some properties which the systems of rational, real, and complex numbers have in common.

27. Conclusion. We have come a long way from the natural numbers we started with. The system of complex numbers did not spring into being full-grown. Natural numbers satisfied the needs of primitive man for an indeterminate length of time. The positive fractions became necessary when people began to share real estate or harvests of grain. These numbers arose in answer to the needs of everyday life. The others, like negative, irrational, and imaginary numbers, arose in answer to the more sophisticated needs of mathematicians. But at each stage we introduced new kinds of numbers because of some inadequacy in the system of numbers already at hand.

Recall that a system of numbers is called closed under a given operation if performing that operation on numbers in the system always yields as a result a number in the system, that is, never takes us outside the system. The system of natural numbers is closed under addition and multiplication since the sum or product of two natural numbers is again a natural number. But the system of natural numbers is not closed under division. We would like our system of numbers to be closed under all our operations, so that we may freely perform the operations and always be sure that the result is in our number system. Therefore we invented the (positive) fractions. The system of positive fractions is closed under addition, multiplication, and division, but not under subtraction. Therefore we invented the negative rational numbers and zero. The entire system of rational numbers is closed under addition, subtraction, multiplication, and division except that division by zero is excluded. But the system of rational numbers is not closed under the operation of taking nth roots. The irrational numbers were introduced because the rational numbers did not provide a number to attach to every point on a line, so that, in particular, not every positive rational num-

ber had a square root among the rational numbers. The system of real numbers is still not closed under the operation of taking nth roots since negative numbers have no square roots among the real numbers. Hence we invented the pure imaginary numbers. The system composed of the real numbers and pure imaginary numbers together would not be closed under addition since the sum of a real number and a pure imaginary number is neither real nor pure imaginary. Hence we invented the system of complex numbers.

The reader may now be in a state of mortal terror lest we have to extend the number system still further if we look at cube roots or some other operation. However, this will not be necessary. The complex number system is not going to be extended any further. There are several reasons for this. One reason is that the complex number system is closed under all the operations to be studied in this book. This will be discussed from a slightly different point of view in the next chapter (section 37). Another reason is that the complex numbers are sufficient for most practical applications. However, there are other number systems, some of which are called hypercomplex number systems, which are studied for various reasons in higher mathematics and which include the complex number system as a special case, much as the complex number system includes the real number system which, in turn, includes the rational number system.



Leopold Kronecker
1823–1891, German

The natural numbers can be considered as the basis for all the numerical portions of mathematics. All the different kinds of numbers introduced can be rigorously defined in terms of the natural numbers, although we have not done this here completely. For example, a fraction was defined as a symbol consisting of a

pair of natural numbers. L. Kronecker (1823–1891) is said to have remarked that "the whole number was created by God, everything else is man's handiwork." The fundamental character of the natural numbers, alluded to by Kronecker, has led not only to the extensive study of their arithmetical properties (the Theory of Numbers,* one of the oldest branches of mathematics) but also, frequently, to the long common practice of attributing to them mystical and supernatural properties (Numerology, one of the oldest branches of balderdash). In fact, the practice of ascribing mystic properties to numbers is not confined to commercial frauds who give marital and financial advice based on numbers assigned to your name, and so on, but has been indulged in by scientists, ancient and modern. Pythagoras, for instance, seems to have been mystically inspired by his discovery that the ratio of the frequencies of harmonious sounds is always expressible in terms of small numbers. For example, if two sounds form an octave, their frequencies are in the ratio of 2/1; if a fifth, their frequencies are in the ratio of 3/2; a fourth, 4/3; a third, 5/4; a minor third, 6/5.† The entire group of Pythagorean disciples mixed mysticism with their science. They felt, for various reasons, that in the whole numbers lay the key to the universe. It may have been this belief that caused them to be so upset about the fact that $\sqrt{2}$ is not expressible as the ratio of two whole numbers that they decided to suppress the information. The human tendency toward mysticism is seen not only in the simultaneous and often intertwined development of the mathematical Theory of Numbers and mystical Numerology but also in the simultaneous and often intertwined development of Astronomy and Astrology. When we are depressed about the various forms of stupidity abroad in the world today, it may be some small comfort to reflect that only three centuries ago as great a man as Kepler wrote with equal seriousness about both astronomy and astrology.

Numerology has been taken with the utmost seriousness by some people, at all times in human history. For example, ac-

* Many arithmetical tricks and amusements depend on the Theory of Numbers. See the chapter on the Theory of Numbers in J. W. A. Young, *Monographs on Topics of Modern Mathematics*, W. W. R. Ball, *Mathematical Recreations and Essays*, and textbooks on the Theory of Numbers.

† Jeans, *Science and Music*, p. 154.

cording to the New Testament,* the number of the beast is 666. This statement was for many years a choice weapon of numerological theologians who would prove a man to be a heretic by attaching the number 666 to his name. This was usually done by assigning numbers to some or all of the letters in some form of his name in some language and showing that these numbers added up to 666. In the middle ages it was no laughing matter to have this done to you. Much the same kind of scheme is used today by the commercial variety of numerologists.

REFERENCES

E. T. Bell, *Numerology*, Williams and Wilkins, Baltimore, 1933.

E. T. Bell, *The Queen of the Sciences*, G. E. Stechert, N. Y., 1938.

T. Dantzig, *Number, the Language of Science*, Macmillan, N. Y., 1939.

A. Dresden, *Invitation to Mathematics*, Holt, N. Y., 1936.

H. B. Fine, *College Algebra*, Part I, Ginn, Boston, 1904.

H. B. Fine, *The Number System of Algebra*, Leach, Shewell, Sanborn, Boston, 1890.

J. H. Jeans, *Science and Music*, Macmillan, London, 1938.

J. W. Young, *Fundamental Concepts of Algebra and Geometry*, Macmillan, N. Y., 1911.

A. N. Whitehead, *Introduction to Mathematics*, Holt, N. Y., 1911.

Histories of Mathematics, as before.

* John, 13:18.

Chapter V

THE LOGIC OF ALGEBRA

28. Algebra as an abstract mathematical science. In the preceding two chapters we sketched the evolution of our complex number system and showed that some of the familiar algebraic manipulations were logical consequences of a few simple definitions and properties, such as the associative, commutative, and distributive laws. In the course of this discussion we found that the number system had certain properties some of which are listed below for reference.

I. *To every pair of numbers a and b, in that order, there is a unique third number denoted by $a + b$, called the sum of a and b.* **(Law of closure for addition.)**

II. $a + b = b + a$. **(Commutative law for addition.)**

III. $(a + b) + c = a + (b + c)$. **(Associative law for addition.)**

IV. *To every pair of numbers a and b, in that order, there is a unique third number denoted by ab or $a \cdot b$, called the product of a and b.* **(Law of closure for multiplication.)**

V. $ab = ba$. **(Commutative law for multiplication.)**

VI. $a(bc) = (ab)c$. **(Associative law for multiplication.)**

VII. $a(b + c) = ab + ac$. **(Distributive law.)**

VIII. *There is a unique number 0, called "zero," such that*

(a) $0 + a = a$, *for any number a;*

(b) $0 \cdot a = 0$, *for any number a;*

(c) *if $a \neq 0$ and $b \neq 0$ then $ab \neq 0$.*

IX. *There is a unique number 1, called "one," such that $1 \cdot a = a$, for any number a.*

X. *If $a < b$ and $b < c$ then $a < c$.*

DEFINITION. *If a number $x > 0$, it is called* ***positive.***
If a number $x < 0$, it is called ***negative.***

XI. (a) *If* $a > 0$ *and* $b > 0$, *then* $ab > 0$.
(b) *If* $a < 0$ *and* $b > 0$, *then* $ab < 0$.
(c) *If* $a < 0$ *and* $b < 0$, *then* $ab > 0$.

XII. $a = a$. **(Law of identity.)**

XIII. *If* $a = b$ *then* $b = a$.

XIV. *If* $a = b$ *and* $b = c$ *then* $a = c$. (*Things equal to the same thing are equal to each other.*)

XV. *If* a *and* b *are any numbers, there exists a unique number* x *such that* $b + x = a$.

DEFINITION. *This number* x *is called* $a - b$. *The number* $0 - b$ *is denoted by* $-b$.

XVI. *If* a *and* b *are any numbers* ($b \neq 0$), *there exists a unique number* x *such that* $bx = a$.

DEFINITION. *This number* x *is called* a/b.

XVII. *If equals are* (a) *added to* (b) *subtracted from* (c) *multiplied by* (d) *divided by equals, the results are equal.*

Properties X and XI apply only to real numbers since the relation $<$ applies only to real numbers.

The principal objective of this chapter will be to explain the logical justification of many more of the familiar manipulations of algebra which you have learned by rote in high school. In particular we shall see that many results of algebra are merely logical consequences of properties I–XVII. There is no need for the student to memorize all these properties since we shall not use many of them formally.

Having justified many of the familiar elementary manipulations in Chapters III, IV, and V, we shall use them freely in the remainder of the book without explicit mention.

Before proceeding to the main business of the chapter, however, let us point out, in passing, how we could make an abstract mathematical science out of algebra. Recall that an abstract mathematical science is constructed by taking some undefined terms and unproved assertions about them (postulates) and then defining new terms and deducing new assertions (theorems). As remarked in Chapter II, section 8, our choice of undefined terms and postulates is usually guided by experience and an abstract mathematical science is usually made with some concrete inter-

pretation or application in mind. Our experience with numbers and their concrete applications to counting and measuring, in Chapters III and IV, suggest the following choice of postulates and undefined terms for an abstract mathematical science called the *algebra of (complex) numbers.*

Let us take the term "number," the operations of "addition," and "multiplication," and the relations "greater than," "less than," and "equal to" as undefined terms, and let statements I–XVII be some of the postulates, all letters representing arbitrary "numbers."

All of elementary algebra, that is the algebra of (complex) numbers, can be deduced logically from such a set of postulates, although several more postulates than we have listed here would be needed to do this completely. However, a completely logical derivation of elementary algebra from such a set of postulates would be too difficult to do here.

For example, from the distributive law, and others, we may deduce immediately, as in Chapter III, such results as

$$a(b + c - d - e + f + \cdots + k) = ab + ac - ad - ae + af + \cdots + ak$$

and

$$(b + c - d - e + f + \cdots + k)a = ba + ca - da - ea + fa + \cdots + ka$$

which we shall use freely and call the **generalized distributive law.** Also as in Chapter III, we shall use the right to insert or remove parentheses in a sum of terms, (or a product of factors) calling the justification for this the **generalized associative law for addition** (or **multiplication**). (See sections 12–15.)

Note that in our present abstract mathematical science, "number," "sum," "product" are completely undefined terms. Thus counting and measuring and all the work of Chapters III and IV may be regarded as concrete interpretations or applications of our present abstract mathematical science. For example, we need not assume the tables of addition on the basis of experience here (as we did in Chapter III) because we can deduce them logically from our postulates. Let us indicate briefly how this may be done. By postulate IX, there is a number 1. By postulate I, there is a number $1 + 1$; call it 2. By postulate I, there is a num-

ber $2 + 1$; call it 3. By postulate I, there is a number $3 + 1$; call it 4; and so on. We can now prove the following important theorems.

THEOREM 1. $2 + 2 = 4$.

Proof. By definition, $2 = 1 + 1$ and $4 = 3 + 1 = (2 + 1) + 1$. We have to prove that $2 + 2 = 4$ or,

$$2 + (1 + 1) = (2 + 1) + 1.$$

But this follows immediately from the associative law for addition.

THEOREM 2. $2 \cdot 2 = 4$.

Proof. By definition, $2 = 1 + 1$. Hence we have to prove that $2(1 + 1) = 4$. But by the distributive law, $2(1 + 1) = 2 \cdot 1 + 2 \cdot 1$. By the commutative law for multiplication and postulate IX, $2 \cdot 1 = 1 \cdot 2 = 2$. Thus $2(1 + 1) = 2 + 2$ which is 4 by Theorem 1. This is what we had to prove.

In a similar fashion we could deduce the rest of the tables of addition and multiplication.

That arithmetic can be done with "number" undefined should not surprise us. The usual *meaning* of the word number is surely not understood by an inanimate calculating machine, but the machine gives us correct answers because it is constructed so as to follow the formal rules by which numbers are combined. Thus arithmetic and algebra can be considered as an abstract mathematical science having no logically necessary connection with counting and measuring, but which does have counting and measuring among its applications or concrete interpretations.

EXERCISES

Defining 5 as $4 + 1$, 6 as $5 + 1$, and so on, prove, by means of our postulates, that:

1. $2 + 3 = 5$. **2.** $2 + 4 = 6$. (Hint: use exercise 1.)

3. $3 + 3 = 6$. (Hint: use exercise 2.)

4. $3 \cdot 2 = 6$. (Hint: use exercise 3.)

5. $6 + 2 = 8$. **6.** $5 + 3 = 8$. **7.** $4 + 4 = 8$. **8.** $2 \cdot 4 = 8$.

29. Algebraic expressions. A letter which may represent different numbers during the same discussion is called a **variable.**

A symbol which is permitted to represent only one certain number throughout a discussion is called a **constant** even if we do not specify what number it represents. It is customary to use the later letters of the alphabet for variables and the earlier ones for constants. For example, x, y, z will usually represent variables and a, b, c, d will represent constants; of course specified numbers like 2, $-3/2$, $\sqrt{2}$, or $3 + 2\sqrt{-4}$ are constants.

The operations of addition, subtraction, multiplication, and division are called **rational operations.** The rational operations together with the operation of taking nth roots, where n is any positive integer, are called **algebraic operations.** An expression which can be built up from the variable x and any constants by means of a finite (limited) number of algebraic operations is called an **algebraic expression in *x*.** For example,

$$(2x + \sqrt{3})\sqrt[3]{3x - 1} + \frac{a}{x^4}$$

is an algebraic expression in x. Note that x^4 is built up by means of multiplication alone; that is, $x^4 = xxxx$.

Remark. As an example of a non-algebraic expression we might mention the "exponential" expression 2^x. This is not algebraic because the exponent x is a variable. On the other hand, the expression x^2 is algebraic since $x^2 = xx$.

An expression which can be built up from the variable x and any constants whatever by means of a finite number of rational operations alone is called a **rational expression in *x*.** Thus

$$\frac{1}{x}, \quad \frac{2x^2 - 3x + 5}{3x - 7}, \quad \frac{1/2}{x - 2} + \frac{x}{x + 3}, \quad \frac{2x^2 + ax - \sqrt{-1}}{\sqrt{3} + x}$$

are rational expressions in x. Note that radicals may occur in a rational expression; the essential thing is that the expression can be built up in such a way that x does not occur under a radical. The last example *is* a rational expression since it may be built up from x and the constants 2, a, $\sqrt{-1}$, $\sqrt{3}$ by rational operations alone. An algebraic expression in x which is not a rational expression (that is, which necessarily involves an nth root of some expression involving x) is called an **irrational expression in *x*.** Thus $1 + \sqrt{x}$ is an irrational expression in x. A rational expression in x which can be built up from the letter x and any constants by means of the operations of addition, subtraction and

multiplication alone (but not division), is called a **polynomial in** x. For example, x, $2x^2 - 7$ and

$$ax^3 + \frac{\sqrt{3}}{2}x^2 - \frac{3}{5}x + (4 + 2\sqrt{-3}) \tag{1}$$

are polynomials in x. Note that the last example is a polynomial in x even though divisions occur; the essential thing is that the expression can be built up in such a way that we do not divide by any expression involving x. For example, (1) can be built up from the letter x and the constants a, $\sqrt{3}/2$, $3/5$, and $4 + 2\sqrt{-3}$ by means of the operations of addition, subtraction, and multiplication alone, and is therefore a polynomial in x. On the other hand, the rational expression $1/x$ is not a polynomial. Polynomials are a special kind of rational expressions, and rational expressions are a special kind of algebraic expressions.

Note that the three terms "rational number," "rational operation," and "rational expression" have entirely different meanings.

Every rational expression in x can be reduced to a quotient of two polynomials in x, by performing the indicated operations. For example, the rational expression $\frac{1}{x} + \frac{5}{x - 3}$ can be reduced to $\frac{6x - 3}{x^2 - 3x}$, by addition of fractions.

The constant factors in each term are called **coefficients.** The coefficients in (1) are a, $\sqrt{3}/2$, $3/5$, and $4 + 2\sqrt{-3}$. The **degree of a polynomial in x** is the highest exponent occurring in a power of x when all indicated operations have been carried out as far as possible. For example, $(x + 1)(x + 2)$ is a polynomial of degree 2 because it becomes $x^2 + 3x + 2$ when multiplied out; similarly $(x + 1)(x + 2) - x^2$ is a polynomial of degree 1 since it reduces to $3x + 2$. Polynomials of degree 1, 2, 3 are called **linear, quadratic,** and **cubic,** respectively. For example, $2x - 3$, $3x^2 + 4x - 1$, and $2x^3 - 5$ are linear, quadratic, and cubic polynomials, respectively.

Similarly we have algebraic expressions, rational and irrational expressions, and polynomials in several variables. For example, $3x^2y^3 + 4x^2y - 3y + x - 3$ is a polynomial in x and y. The **degree of a term involving more than one variable** is defined to

mean the sum of the exponents attached to these variables. The **degree of a polynomial in more than one variable** is defined to mean the highest of the degrees of its terms, after all indicated operations have been carried out as far as possible. The degree of the polynomial in the last example is 5.

Algebraic expressions do not represent definite numbers because of the presence of variables. If, however, we replace the variable by a definite number, the expression is either undefined or takes on a definite value. For example, the rational expression $(12 - x)/x$ has the value 5 if we let $x = 2$, the value 3 if we let $x = 3$, the value 7/5 if we let $x = 5$, the value 0 if we let $x = 12$, and no value at all if we let $x = 0$ because division by zero is excluded. Note that to say an expression has no value at all is not the same as to say it has the value zero.

EXERCISES

(*a*) *List all the terms "algebraic expression," "irrational expression," "rational expression," "polynomial" which apply to each of the following;*

(*b*) *Find the value of each expression when $x = 3$, $y = 2$:*

1. $\sqrt{3x + 7}$. **2.** $6/x$. **3.** $\frac{1}{2}x^2 + \sqrt{3}$. **4.** $\dfrac{5x^2 - 7x + 1}{3 - x^3}$.

5. x. **6.** $x + 3y$. **7.** $\dfrac{2x - 3}{x + 7}$. **8.** $\dfrac{\sqrt{x} - 2}{2x + 1}$.

9. $2x^3 + 4x - 7$. **10.** $\sqrt{x^2 + 16}$.

11. If rational numbers are substituted for the variables in a rational expression whose coefficients are rational numbers, what kind of number must the resulting value of the expression be? Explain and illustrate.

12. If integers are substituted for the variables in a polynomial whose coefficients are integers, what kind of number must the resulting value of the expression be? Explain and illustrate.

13. Will the value of an irrational expression for any value of the variable always be an irrational number? Explain and illustrate.

30. Polynomials. Addition and multiplication. To add two polynomials, for example, $2x^2 + 3x + 5$ and $3x^2 + 4x - 3$, you have learned to write "like powers" beneath each other and proceed as follows:

$$\begin{array}{r} 2x^2 + 3x + 5 \\ 3x^2 + 4x - 3 \\ \hline 5x^2 + 7x + 2. \end{array}$$

Let us see how this result follows logically from our postulates. We are asked for the sum of the two expressions

$$(2\,x^2 + 3\,x + 5) + (3\,x^2 + 4\,x - 3).$$

For any value of x, the individual terms may be regarded as single numbers, because of the law of closure for multiplication. By the generalized associative law for addition we may remove the parentheses. Then by the commutative law for addition we may rearrange the order of the terms, obtaining

$$2\,x^2 + 3\,x^2 + 3\,x + 4\,x + 5 - 3.$$

By the generalized associative law for addition we may regroup the terms like this:

$$(2\,x^2 + 3\,x^2) + (3\,x + 4\,x) + (5 - 3).$$

By the generalized distributive law $2\,x^2 + 3\,x^2 = (2 + 3)x^2$ and $3\,x + 4\,x = (3 + 4)x$. Thus we obtain

$$5\,x^2 + 7\,x + 2,$$

by substitution.

Let us recall the fact that $x^2 \cdot x^3 = x^5$. This follows from the definition of exponent. For $x^2 = xx$ and $x^3 = xxx$. Hence $x^2x^3 = (xx)(xxx) = xxxxx = x^5$. In general, we have the following theorem.

THEOREM 1. *If m and n are any positive integers, $x^m \cdot x^n = x^{m+n}$.*

Proof. By definition $x^m = xx \cdots x(m$ factors) and $x^n = xx \cdots x(n$ factors). Hence

$$x^m \cdot x^n = (\overbrace{xx \cdots x}^{m \text{ factors}}) \cdot (\overbrace{xx \cdots x}^{n \text{ factors}}) = xx \cdots x(m + n \text{ factors}).$$

Hence $x^m \cdot x^n = x^{m+n}$.

We can now recall the process you learned for multiplying polynomials, for example, $2\,x + 3$ and $3\,x + 4$:

$$\begin{array}{r} 2\,x + 3 \\ 3\,x + 4 \\ \hline 8\,x + 12 \\ 6\,x^2 + 9\,x \phantom{{}+12} \\ \hline 6\,x^2 + 17\,x + 12. \end{array}$$

How can this result be justified on the basis of our postulates? We want the product $(2\,x + 3)(3\,x + 4)$. By the laws of closure

for addition and multiplication, $2x + 3$ is a single number for any particular value of x. Hence, treating the first parenthesis as a single quantity with a long name, the distributive law enables us to write

$$(2x + 3)(3x + 4) = (2x + 3)3x + (2x + 3)4.$$

Using the generalized distributive law on each of the terms of the right member, we get

$$2x \cdot 3x + 3 \cdot 3x + 2x \cdot 4 + 3 \cdot 4.$$

By the commutative law for multiplication we obtain

$$2 \cdot 3 \cdot x \cdot x + 3 \cdot 3 \cdot x + 2 \cdot 4 \cdot x + 3 \cdot 4.$$

By the generalized associative law for multiplication we get

$$(2 \cdot 3)(x \cdot x) + (3 \cdot 3)x + (2 \cdot 4) \cdot x + 3 \cdot 4$$

or

$$6x^2 + 9x + 8x + 12.$$

By the generalized associative law for addition this may be written

$$6x^2 + (9x + 8x) + 12.$$

But $(9x + 8x) = (9 + 8)x = 17x$ by the generalized distributive law. Hence we obtain, finally, $6x^2 + 17x + 12$.

Of course the short arrangement of the work you learned in school is far quicker than what we have done and no one in his right mind would prefer what we have done *if his purpose was to obtain the right answer quickly.* But our purpose is not technical facility but the logical deduction of our results from our postulates or assumptions. We have thus indicated that the queer processes of addition and multiplication of polynomials, which you learned by memory in school, may be justified on the basis of our postulates. We shall not stop to show how more complicated processes like division or extraction of square roots may be justified similarly on the basis of our postulates although this can be done.

EXERCISES

Perform the indicated operations by means of the usual short scheme, and then derive the same result from our postulates, justifying each step:

1. Add $2x^2 + 3x + 5$ and $3x^2 + 2x + 1$.
2. Multiply $4x + 2$ by $3x + 5$.

3. Add $6x^3 + 2x + 6$ and $x^3 + 2x^2 + 3x + 1$.
4. Multiply $x^2 + 2x + 3$ by $3x + 1$.
5. Multiply $(x + 2)(x + 3)$.
6. Multiply $(x + y)(x + y)$.
7. Multiply $(2ax + b)(2ax + b)$.
8. Multiply $(x + y)(x - y)$.
9. Multiply $(x - y)(x - y)$.
10. Multiply $7x + 3$ by $3x - 1$.
11. Add $2x + 3xy + y^2$ and $5x - xy - 4y^2$.
12. Multiply $x - y$ by $x^2 + xy + y^2$.
13. Multiply $(x + y)(x^2 - xy + y^2)$.
14. Subtract $x^2 - 3x - 5$ from $3x^2 - x + 2$.

31. Factoring. We have just multiplied two polynomials to obtain their product. Frequently it is convenient to be able to begin with the product and tell what factors were multiplied together to obtain it. This is called **factoring.** It is usually much harder to factor or "unmultiply" than it is to multiply, just as it is much harder to unscramble eggs than it is to scramble them. In fact we shall not take up any systematic method of factoring, but will consider only the simplest special cases.

Case I. Taking out a common factor. For example, $3x^2 + 6xy + 9x^2y = 3x(x + 2y + 3xy)$. This, we have already observed, is merely an application of the (generalized) distributive law.

Case II. Difference of two squares. We observe and remember that $(a + b)(a - b) = a^2 - b^2$ by direct multiplication. Hence we factor $x^2 - y^2 = (x + y)(x - y)$.

Case III. Quadratic trinomials. These are factored by trial and error. For example, to factor $x^2 - 5x + 6$, we see that, if it can be factored, it must be the product of two linear factors of the form $(x - ?)(x - ?)$. We experiment with various values in place of the question marks until we arrive at $(x - 3)(x - 2)$.

In short, *we factor essentially by remembering our experiences in multiplying.* To check the correctness of our factoring, we have only to multiply our factors together and see if we get back the original expression. Whether a polynomial can be factored or not is not a simple question. It depends (among other things) on what kind of numbers you are willing to allow as coefficients in the factors. For example, $x^2 - 2$ is not factorable if you insist on rational coefficients in the factors, but can be factored into

$(x + \sqrt{2})(x - \sqrt{2})$ if you allow irrational coefficients. Similarly, $x^2 + 1$ cannot be factored if you insist on real coefficients in the factors but can be factored into $(x + \sqrt{-1})(x - \sqrt{-1})$ if you allow imaginary coefficients.

EXERCISES

Factor:

1. $4x^2y + 6xy^2 - 10x^3y^3$.
2. $16x^2 - 4a^2b^2$.
3. $x^2 - 7x + 10$.
4. $3x^2 - 8x + 5$.
5. $x^2 - 9$.
6. $(x + 1)^2 - (y + 3)^2$.
7. $2x^2 + 5x + 2$.
8. $3x^2 - x^3 - 10x^4$.
9. $y^2 + 2ky + k^2$.
10. $y^2 - 2ky + k^2$.
11. $4a^2x^2 + 4abx + b^2$.
12. $2x^2 - 5x - 3$.

32. Equations. The equation

$$2(x + 3) = 2x + 6 \tag{1}$$

seems to say that two different expressions are "equal." But an algebraic expression has no definite value unless we first assign a numerical value to the variable x. What, then, does it mean to say that these expressions are "equal"? In this example, it means that no matter what number x represents, the resulting values of the two expressions will be equal. Thus, if $x = 1$, the statement (1) asserts that $2(1 + 3) = 2 \cdot 1 + 6$ or $2 \cdot 4 = 2 + 6$; if $x = 2$, it says that $2(2 + 3) = 2 \cdot 2 + 6$ or $2 \cdot 5 = 4 + 6$. And so on.

DEFINITION. *If an equation becomes a true statement for every value of x for which each expression involved has a value it is called an **identical equation** or an **identity**.*

Thus (1) is an identity. Also,

$$\frac{1}{x-2} + \frac{1}{x-1} = \frac{2x-3}{(x-2)(x-1)}$$

is an identity; it has no sense for the values $x = 1$ and $x = 2$ since division by zero is meaningless, but for all other values of x it is true, as we may verify by adding the two fractions on the left. Similarly, $\frac{1}{x} = \frac{3}{3x}$ is an identity since it is true for all values of x except $x = 0$ for which value the expressions involved have no meaning.

Consider, however, the equation

$$(2) \qquad 2x + 3 = x + 6.$$

This is clearly not an identity, because for $x = 1$ it says that $5 = 7$ which is notoriously false. In fact, an equation cannot be considered an assertion (that is, a proposition) at all until a value has been assigned to x; and when values or meanings have been assigned to the variable x, the resulting statement may be true or false, or simply nonsense. Such statements which contain variables are often called "propositional functions"; they become "propositions" only when we have substituted meanings or values for the variables which make the resulting statements either true or false (see Remark 1, section 5). For example, $x + y = 5$ is a propositional function, while $2 + 3 = 5$ is a true proposition and $2 + 4 = 5$ is a false proposition. Similarly, the equation (2), which is a sentence with the verb "is equal to," is a propositional function. If we substitute $x = 1$ in it, we get the false proposition $5 = 7$. If we substitute $x = 3$ in it, we get the true proposition $9 = 9$. We say $x = 3$ *satisfies the equation* $2x + 3 = x + 6$, or *makes it true;* 3 is called a *root* of the equation. If an equation is not an identity it is called *conditional.* That is, we make the following definition.

DEFINITION 2. *If an equation becomes a false statement for some value of the unknown, it is called a* **conditional equation** *or simply an* **equation.** *A value of x which does satisfy the equation, that is, a value of x for which the equation becomes a true statement, is called a* **root** *of the equation. To* **solve** *an equation means to find its roots.*

A conditional equation may not have any roots; for example, $1/x = 0$, or $x = x + 3$. These are propositional functions which are true for no value of the unknown. Similarly, "x is a man more than 30 ft. tall" is a propositional function which is false no matter what meaning or value we substitute for x.

For example, let us prove that

$$x(x + 3) + 6 = x^2 + 3(x + 2)$$

is an identity. To prove this we try to reduce the statement to something which we recognize as an identity. Using the distributive law we obtain

$$x^2 + 3x + 6 = x^2 + 3x + 6$$

which is clearly true by identity. But what have we really proved? Our reasoning has been: if the first line is true then the second line must be true; and the second line is true. Does it follow that the first line is therefore true? Certainly not! From "A implies B and B is true" it does not follow that A is true (see section 4). What we really need is the converse proposition "if the second line is true then the first must be true," because "B implies A and B is true" *does* imply that A is true. Since the converse proposition does not follow automatically (a valid proposition may have an invalid converse), we must prove it independently. This can be done by simply starting from the bottom and working up to the original statement, *provided* each step we made on the way down is *reversible*. In the above example this is so. In fact, *if the only steps taken are merely substitutions then, clearly, the argument can be reversed.* For if a step was made by replacing a by its equal b, the reverse step can be made by replacing b by its equal a.

But one must guard against non-reversible steps. For in-instance, assuming that $\quad 3 = 7,$
we can write $\quad 7 = 3,$
and therefore, adding both sides of these equations, we obtain

$$10 = 10$$

because if equals are added to equals the results are equal. We cannot conclude from the truth of our conclusion that our hypothesis was true. Similarly the proposition "if $a = b$ then $a^2 = b^2$" is true because if equals are multiplied by equals the results are equal. But the converse "if $a^2 = b^2$ then $a = b$" is false. For instance, from the assumption $-3 = +3$ we can conclude that $9 = 9$ since $(-3)^2 = 3^2$; but from the truth of the latter statement we cannot infer the truth of the former. However, it is true that "if $a^2 = b^2$ then $a = +b$ or $a = -b$." Similarly, from the assumption $3 = 7$ and the fact that $0 = 0$, we get, by multiplication, the result $0 = 0$ since if equals are multiplied by equals the results are equal. But we cannot infer from the truth of the conclusion that the hypothesis was true. All of these examples exhibit non-reversible steps.

A method for verifying identities which does not involve us with the question of reversible steps is to leave one member of

the equation untouched and to substitute various equivalent expressions for the other until it becomes identical in appearance with the first member. This establishes the identity by virtue of the axiom "things equal to the same thing are equal to each other." This procedure is often convenient.

We shall be concerned principally with polynomial equations, that is, equations of which both members are polynomials. Thus,

$$3x^2 + 5x - 6 = x^2 + x - 1$$

is a polynomial equation. By virtue of Postulate XVII, section 28, on equality we may subtract the right member from both sides, obtaining

$$3x^2 + 5x - 6 - (x^2 + x - 1) = x^2 + x - 1 - (x^2 + x - 1),$$

or, using the associative, commutative, and distributive laws, and the rule of signs,

$$2x^2 + 4x - 5 = 0.$$

The process of subtracting the right member from both sides was known to you in school as "transposing." Similarly *every polynomial equation can be written with a polynomial on the left and zero on the right of the equals sign.* When this is done, the degree of the polynomial on the left is called the **degree of the equation.** Equations of degree 1, 2, 3 are called **linear, quadratic,** and **cubic,** respectively. Every linear equation can be written in the form $ax + b = 0$ where a and b are constants ($a \neq 0$). Similarly, every quadratic equation can be written in the form $ax^2 + bx + c = 0$ ($a \neq 0$), every cubic equation can be written in the form $ax^3 + bx^2 + cx + d = 0$ ($a \neq 0$), and so on.

Many elementary practical problems in geometry, surveying, commerce, etc., lead to the necessity for solving equations. Some easy ones are given in section 41. In Europe these problems began to demand attention in the 13th century, when algebra, as we know it, began its slow development. We shall now study the problem of solving or finding the roots of equations.

EXERCISES

Verify the identities:

1. $x^2 + 2(x - 3) = x(x + 2) - 6.$

2. $\dfrac{2}{x - 1} + \dfrac{1}{x} = \dfrac{3x - 1}{x^2 - x}.$

3. $\dfrac{1}{x} + \dfrac{2}{x - 3} = \dfrac{3x - 3}{x^2 - 3x}.$

4. $\dfrac{2x}{3x-1} + \dfrac{2}{x-2} = \dfrac{2(x^2+x-1)}{(3x-1)(x-2)}$.

5. $2(3x + [x+1] + 2[x + 3\{x-1\}]) = 24x - 10$.

6. $\dfrac{x+2}{x} + \dfrac{x+3}{3x} = \dfrac{4x+9}{3x}$.

7. $\dfrac{\frac{a}{b} + \frac{b}{c}}{\frac{a}{b} - \frac{b}{c}} = \dfrac{ac+b^2}{ac-b^2}$.

8. $\dfrac{5x}{x-2} - \dfrac{3}{2} = \dfrac{7x+6}{2(x-2)}$.

Find the degree of each of the following equations:

9. $3x^4 + 2x^3 - x^2 + 3 = 3x^4 - x^3 + 2x - 1$.

10. $x^3 - x + 1 = x^3 - 3x + 4$.

11. $2x^4 + 3x^3 - x^2 + 2 = 2x^4 + 3x^3 - 3x^2 - x - 5$.

Decide whether each of the following equations is an identity or a conditional equation:

12. $3(x-2) = 3x - 6$.

13. $5(x - \frac{1}{15}) = 5x - 3$.

14. $\dfrac{1}{x} + \dfrac{2}{2x} = \dfrac{3}{3x}$.

15. $\sqrt{x^2+9} = x + 3$.

16. $5 - 2x = 3x$.

17. $\dfrac{5x}{3} - \dfrac{2x}{3} = x$.

33. Linear equations. The procedure which some students employ to solve an equation like $3x + 5 = x + 11$ is something like this. They first "transpose" the 5 and the x being careful to change the signs of transposed terms (because they were taught to do so), obtaining $3x - x = 11 - 5$ or $2x = 6$. Then they "bring the 2 across to the other side" and place it under the 6 being careful *not* to change the sign of the 2 (because they were taught to do this), obtaining $x = 6/2$ or $x = 3$. After the 3 they write "answer" as a kind of solemn "amen" to the entire ritual.

Let us examine carefully the logic of this process. We wish to find a root of the equation, provided it has any roots—which is something we do not know in advance. Suppose x is a root, that is, a number which satisfies the equation

$$(1) \qquad 3x + 5 = x + 11.$$

We would like to obtain an equation with no x's in the right member and no constant term in the left. This suggests subtracting x from both sides to remove the x from the right member of (1) and subtracting 5 from both sides to remove the 5 from the

left side of (1). Therefore we reason as follows. If the number x satisfies (1) then it also satisfies

$$3x + 5 - (x + 5) = x + 11 - (x + 5)$$

since we may subtract $(x + 5)$ from both sides by virtue of the axiom "if equals are subtracted from equals, the results are equal." By means of the associative, commutative, and distributive laws, we obtain

$$(2) \qquad 2x = 6.$$

But if x is a number such that $2x = 6$, we may divide both sides of (2) by 2 by virtue of the axiom "if equals are divided by equals, the results are equal." Thus we obtain the conclusion $x = 3$. But what have we proved as a result of this chain of reasoning? Clearly we have proved the proposition "*if* x is a number satisfying $3x + 5 = x + 11$ *then* $x = 3$." Can we then assert that 3 *does* satisfy the original equation? Clearly not! For to say this is to assert the proposition "*if* $x = 3$ *then* x satisfies the equation $3x + 5 = x + 11$." This is, however, the *converse* of what we proved and we know that a proposition may well be valid without having a valid converse. All we have really proved is that if x is a root of the original equation then x cannot be anything else but 3. That is, 3 is the only possible candidate eligible for the position of root. But, so far as we know, the equation may have no roots at all! What can we do to see if 3 is a root or not? One procedure would be to substitute 3 for x in the equation $3x + 5 = x + 11$ and verify directly whether or not it satisfies the equation. Another satisfactory procedure would be to prove the converse proposition by reasoning backwards from the last step to the first, which can be done provided all the steps taken originally are reversible steps. This is so in the case of linear equations because all we use is axiom XVII, section 28, and if to get from one step to the next we add the same thing to both sides we can get back again by simply subtracting the same thing from both sides, and so on. If we never multiply by zero we will have no trouble with division by zero. This is why there is no need to worry about the converse proposition in the case of linear equations. The following example, which is not itself linear, but which leads to a linear equation, shows that we cannot always be so carefree.

Example. Solve the equation

$$\frac{2}{x} + \frac{x+2}{x(x-2)} = \frac{4}{x(x-2)}.$$

The equation has no sense for $x = 0$ and $x = 2$. (Why?) We may multiply both sides of the equation by the common denominator $x(x - 2)$, since if equals are multiplied by equals the results are equal, obtaining

$$x(x-2)\left[\frac{2}{x} + \frac{x+2}{x(x-2)}\right] = \frac{4}{x(x-2)} \cdot x(x-2).$$

Using the distributive law in the left member, we get

$$x(x-2) \cdot \frac{2}{x} + x(x-2) \cdot \frac{(x+2)}{x(x-2)} = \frac{4}{x(x-2)} \cdot x(x-2).$$

Simplifying the fractions, this becomes

$$2(x-2)+(x+2) = 4,$$

except, perhaps, for $x = 0$ or $x = 2$, in which cases the cancellation is not valid. (Why?) Applying the generalized distributive law, we have

$$2x - 4 + x + 2 = 4.$$

By the commutative, associative, and distributive laws, etc., this becomes

$$3x - 2 = 4.$$

Or, $3x = 6$ (Reason?)

or, $x = 2.$ (Reason?)

But $x = 2$ is not a root of the original equation, since substituting $x = 2$ causes the term $\frac{4}{x(x-2)}$ to become $\frac{4}{0}$ which is meaningless. All that our reasoning above established was that *if* the original equation had a root it would have to be 2. But 2 does *not* satisfy the equation. Hence this equation has no roots. *This indicates the necessity for establishing the converse proposition (most easily done by direct substitution in the original equation) before asserting that your "answer" is really a root of the equation.*

EXERCISES

Solve, justifying each step on the basis of our postulates:

1. $3x - 5 = 7x + 8.$

2. $7 - 6x = 3x - 12.$

3. $\frac{1}{x} + \frac{3}{x} = \frac{1}{3}$.

4. $6x - 2(x - 3) = x + 8$.

5. $7x - 3 = 2(x + 3) - 4$.

6. $\frac{1}{6} + \frac{1}{10} = \frac{1}{x}$.

7. $\frac{2}{x - 1} + \frac{1}{(x - 1)(x - 3)} = \frac{1}{(x - 1)(x - 3)}$.

34. Solution of quadratic equations by factoring. Some students learn in school to solve the equation

$$x^2 - 5x + 6 = 0 \tag{1}$$

as follows. They factor the left side, obtaining

$$(x - 2)(x - 3) = 0. \tag{2}$$

Then they draw a T and tear the equation in half, making two equations out of it, completing the ritual thus:

$x - 2 = 0$	$x - 3 = 0$
$x = 2$	$x = 3$
ans.	ans.

What is the reasoning behind this mysterious procedure? In particular, what logical justification is there for ruthlessly splitting the equation into two equations? We say, as before, *if* x is a number such that (1) is true (assuming there *are* any such numbers), then x must make (2) true as well, since the left members of (1) and (2) are equal *identically* (that is for all values of x). But by postulate VIII (c) of section 28, ***the product of two quantities can be 0 only when one or the other (or both) of the two quantities is itself zero.*** (See also theorem 1, section 22.) But the first parenthesis is zero only when $x = 2$, and the second is zero only when $x = 3$. We have proved that "*if* x is a number such that $x^2 - 5x + 6 = 0$ *then* x can only be either 2 or 3." That is, 2 and 3 are the only eligible candidates for the rootship. Before we can say that they are really roots, we need the converse propositions, "*if* $x = 2$ *then* x satisfies the equation $x^2 - 5x + 6 = 0$," and "*if* $x = 3$ *then* x satisfies the equation $x^2 - 5x + 6 = 0$." As in the last section, this can be proved either by direct substitution in (1) or by proceeding from the bottom line to the top by reversing each step of reasoning. We leave this to the reader as an exercise.

Note that the following "solution" is *incorrect:*

$$x^2 - 5x + 6 = 12$$
$$(x-2)(x-3) = 12$$

$x - 2 = 12$	$x - 3 = 12$
$x = 14$	$x = 15.$

The third line is erroneous because we cannot make for the number 12 a statement like the statement in heavy type above for the number 0. That is, it is not true that the product of two quantities can be 12 only when one or the other (or both) of the quantities is itself 12. Zero is the only number with this property.

EXERCISES

Solve, justifying each step:

1. $x^2 - 7x + 10 = 0.$
2. $x^2 - 5x + 6 = 20.$
3. $3x^2 - 11x - 20 = 0.$
4. $(x+3)(2x-1) = 15.$
5. $x^2 - 9 = 0.$
6. $(x-4)(x-1) = -2.$
7. $(x+5)(x+1) = 12.$
8. $6x^2 + 7x - 3 = 0.$
9. $4x^2 - 25 = 0.$
10. $9x^2 - 1 = 0.$
11. $\dfrac{x^2}{(x-2)(x-3)} = \dfrac{2}{x-2} + \dfrac{6}{(x-2)(x-3)}.$

35. Irrational equations. We shall see again that our fussing about the converse proposition in the two preceding sections was not mere academic purism.

Example 1. Solve

(1) $$x - 7 = \sqrt{x-5}.$$

This is called an **irrational equation** because while both sides are algebraic expressions, at least one of them is not a rational expression. Recall that we have agreed that $\sqrt{\ }$ shall mean the *positive* square root, wherever possible, to avoid ambiguity. To solve (1), we reason as follows. If x is a number satisfying (1), (provided there *are* any such numbers), then, squaring both sides, $x^2 - 14x + 49 = x - 5$, because if $a = b$ then $a^2 = b^2$. Then $x^2 - 15x + 54 = 0$, or $(x-6)(x-9) = 0$. Hence $x = 6$ or $x = 9$. We have proved that "*if* x satisfies (1), *then* $x = 6$ or $x = 9$." What about the converse proposition? Do $x = 6$ and $x = 9$ really satisfy (1)? We have only proved so far that they are the only *possible* roots, that is, the only candidates eligible

for the position of root. Let us see whether or not they *are* roots by substituting in (1). Substituting $x = 6$, we obtain $6 - 7 = \sqrt{6-5}$ or $-1 = 1$. Thus 6 is *not* a root. On the other hand, substituting $x = 9$, we get $9 - 7 = \sqrt{9-5}$ or $2 = 2$, so that 9 *is* a root.

Example 2. Solve

$$(2) \qquad \sqrt{x-2} = \sqrt{x} + 2.$$

Squaring both sides we get $x - 2 = x + 4\sqrt{x} + 4$, or $-6 = 4\sqrt{x}$, or $-3 = 2\sqrt{x}$. Squaring again, we get $9 = 4x$, or $x = 9/4$. Therefore 9/4 is the only possible root. Substituting in (2), we get

$$\sqrt{\frac{9}{4} - 2} = \sqrt{\frac{9}{4}} + 2, \text{ or } \sqrt{\frac{1}{4}} = \sqrt{\frac{9}{4}} + 2, \text{ or } \frac{1}{2} = \frac{3}{2} + 2,$$

which is clearly not true. Hence 9/4 is not a root, and our equation has no roots at all.

The unsuccessful candidates for the position of root, which turn up sometimes in these examples, are often called *extraneous roots*, but this is merely a euphemistic way of saying that they are not roots at all. Of course, not all irrational equations have extraneous roots.

The occurrence of extraneous roots shows clearly that substituting our possible answers in the original equation is more than a mere (superfluous) "check" but is really an essential part of the argument.

Since extraneous roots occur in the above examples, it must be that the converse proposition is not true and hence that not every step in the reasoning on the way down is reversible. Clearly the non-reversible step is the step of squaring. For although "if $a = b$ then $a^2 = b^2$" is valid, the converse "if $a^2 = b^2$ then $a = b$" is not, as we saw in section 27.

EXERCISES

Solve:

1. $\sqrt{x+2} = 3$.

2. $\sqrt{x+2} = -3$.

3. $\sqrt{x} = x - 2$.

4. $\sqrt{x} = 2 - x$.

5. $\sqrt{3x+4} = 2 + \sqrt{2x-4}$.

6. $\sqrt{x-4} = 9 - \sqrt{x+5}$.

7. $\sqrt{x-4} = \sqrt{x+5} - 9$.

8. $\sqrt{x+4} + \sqrt{x+11} = 7$.

9. $\sqrt{x+2}+\sqrt{2x+5}=1.$

10. $\sqrt{3x+1}=1+\sqrt{2x-1}.$

11. $\sqrt{2x+4}=\sqrt{2x}+1.$

12. $\sqrt{3x-1}+\sqrt{3x+6}=7.$

13. $\sqrt{3-2x}=3+\sqrt{2+2x}.$

36. Solution of quadratic equations by formula. We have seen that every quadratic equation can be written in the form

$$(1) \qquad ax^2 + bx + c = 0$$

where $a \neq 0$. (If $a = 0$, the equation is not really quadratic.) We shall now derive the familiar formula for the roots of any quadratic equation. It is not likely that the reader would think of making the first few steps himself, except after much experimentation. Their purpose is to obtain a perfect square on the left side of the equation.* But the reader can and should supply the justification for each step. Subtracting c from both sides of (1), we obtain

$$ax^2 + bx = -c.$$

Multiplying both sides by $4a$, we have

$$4a^2x^2 + 4abx = -4ac.$$

Adding b^2 to both sides, we get

$$4a^2x^2 + 4abx + b^2 = b^2 - 4ac.$$

The left member can be factored (see exercise 11, section 31), obtaining

$$(2ax + b)^2 = b^2 - 4ac.$$

Taking the square root of both sides, we have

$$2ax + b = \pm\sqrt{b^2 - 4ac},$$

since if $u^2 = v^2$ then $u = \pm v$. Hence,

$$2ax = -b \pm \sqrt{b^2 - 4ac}.$$

Since $a \neq 0$ and hence $2a \neq 0$, we may divide both sides by $2a$, obtaining

$$(2) \qquad x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

We have proved that "*if x is a root of* $ax^2 + bx + c = 0$ *then*

* For a discussion of the technical device of "completing the square," see the Appendix, section 168.

$x = \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$." We leave it to the reader to show that the converse proposition is also true, either by verifying that each step is reversible no matter which of the answers we start with, or by substitution.

We can apply our formula to solve any quadratic by merely substituting the values of a, b, and c in (2).

Example 1. Consider the equation $x^2 - 5x + 6 = 0$. Here $a = 1$, $b = -5$, $c = 6$. Hence the roots are

$$\frac{-(-5) \pm \sqrt{(-5)^2 - 4 \cdot 1 \cdot 6}}{2 \cdot 1} = \frac{5 \pm 1}{2} = \begin{cases} 3 \\ 2. \end{cases}$$

Example 2. Consider the equation $x^2 + 2x + 4 = 0$. Here $a = 1$, $b = 2$, $c = 4$. Hence the roots are

$$\frac{-2 \pm \sqrt{2^2 - 4 \cdot 1 \cdot 4}}{2 \cdot 1} = \frac{-2 \pm \sqrt{-12}}{2} = \frac{-2 \pm \sqrt{(-3)4}}{2}$$

$$= \frac{-2 \pm 2\sqrt{-3}}{2} = \frac{2(-1 \pm \sqrt{-3})}{2} = \begin{cases} -1 + \sqrt{-3} \\ -1 - \sqrt{-3}. \end{cases}$$

Notice that the roots of a quadratic equation may be imaginary, even though the coefficients are real.

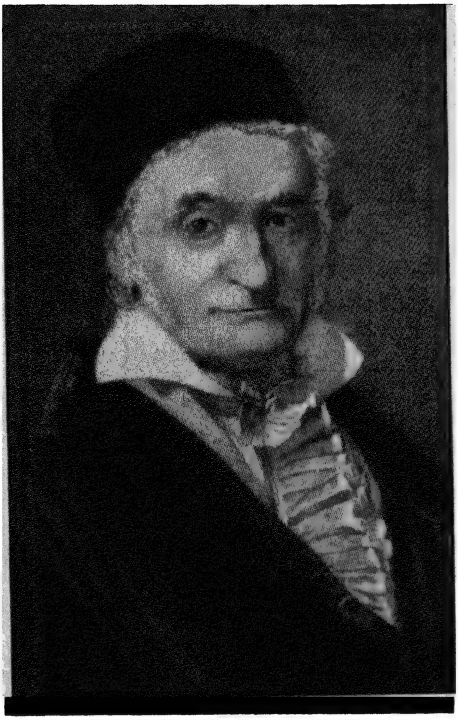
EXERCISES

Solve by formula and tell whether the roots are imaginary, real, rational, or irrational:

1. $x^2 - 5x + 6 = 0$.	**2.** $x^2 - 9 = 0$.	**3.** $x^2 - 5x + 1 = 0$.
4. $2x^2 - 5x - 3 = 0$.	**5.** $x^2 + 9 = 0$.	**6.** $x^2 + x + 3 = 0$.
7. $3x^2 - 11x = 20$.	**8.** $6x^2 = 3 - 7x$.	**9.** $(x + 5)(x + 1) = 13$.
10. $3x^2 + x + 5 = 0$.	**11.** $x^2 + 7x + 1 = 0$.	**12.** $x^2 + 16 = 0$.

37. The fundamental theorem of algebra. We can now review briefly the growth of the number system in the light of the theory of equations and see why we made the statement, at the close of the preceding chapter, that the complex number system need not be enlarged further. Suppose we decide that we want all polynomial equations to have roots. Now let us imagine that we have no numbers in our possession except the natural numbers. Then a simple linear equation like $2x = 3$ has no root.

In order to remedy this condition, we invent fractions. But a simple linear equation like $x + 5 = 2$ has no root even among the fractions. Hence we invent negative numbers. A simple quadratic equation like $x^2 = 2$ has no root among all the (positive and negative) rational numbers, as we proved in section 24. Therefore we invent the irrational numbers which together with the rational numbers complete the system of real numbers. However, a simple quadratic equation like $x^2 = -1$ has no root among all the real numbers (section 26). Hence we invent the pure imaginary numbers. But a simple quadratic equation like $x^2 + 2x + 4 = 0$ has no roots among either the real or pure imaginary numbers (example 2, section 36). Therefore we invent the complex numbers. Now we might well expect that there might be some equation of degree 3 or higher which has no roots even in the entire system of complex numbers. That this is not the case was known to Karl Friedrich Gauss, who proved (in 1799) the following theorem, the truth of which had long been suspected.



Karl Friedrich Gauss
1777–1855, German

Fundamental Theorem of Algebra. *Every polynomial equation, no matter how high its degree, with coefficients in the complex number system, has a root among the complex numbers.*

For example, an equation like

$$(2 + 3\sqrt{-5})x^{1941} + \frac{\sqrt{2}}{3}x^{1776} + \frac{1865}{7}x^{1492} + \sqrt{-11}\,x + 13 = 0$$

is known in advance to have a root among the complex numbers. The proof of this theorem is too difficult to be taken up here. We

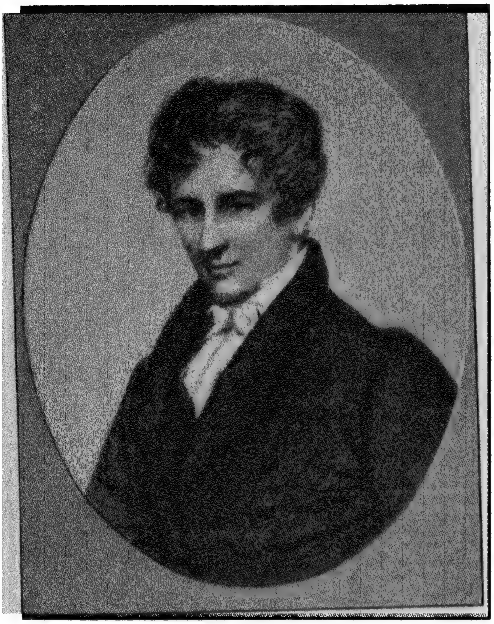
can see, however, that for the purpose of solving polynomial equations we do not need to extend the number system any further.

38. Algebraic formulas for the roots. The general linear equation can be written in the form $ax + b = 0 (a \neq 0)$. Hence a formula for its roots is $x = \dfrac{-b}{a}$. Every quadratic equation can be written in the form $ax^2 + bx + c = 0 (a \neq 0)$. Its roots, as we have seen, are given by the formulas $\dfrac{-b + \sqrt{b^2 - 4ac}}{2a}$ and $\dfrac{-b - \sqrt{b^2 - 4ac}}{2a}$. Both the formulas for the roots of the general linear equation $ax + b = 0$ and the roots of the general quadratic equation $ax^2 + bx + c = 0$ give the roots as algebraic expressions * in the coefficients. The mathematician's desire for general results makes it natural to ask the following question. Can we get similar formulas giving the roots as algebraic expressions in terms of the coefficients for the general equation of any degree? For the general cubic equation $ax^3 + bx^2 + cx + d = 0$, such formulas were substantially obtained by Tartaglia (about 1540), though they are often referred to as Cardan's formulas because they were first published in 1545 by Cardan, who obtained them from Tartaglia under a pledge of secrecy.† For the general equation of degree 4, $ax^4 + bx^3 + cx^2 + dx + e = 0$, such formulas were obtained by Ferrari at about the same time. We shall not state the formulas for the general equations of degree 3 and 4 here. The next task was, naturally, to obtain similar formulas for the general equation of degree five, $ax^5 + bx^4 + cx^3 + dx^2 + ex + f = 0$. Attempts to find such formulas were made from the 16th century until early in the 19th century without success. The reason for this failure became evident in 1824 when N. H. Abel, a brilliant young Norwegian mathematician, proved, at the age of 22, that it is not possible to write the roots of the general equation of degree higher than four as algebraic expressions in terms of the coefficients.

* Recall the definition of "algebraic expression" (section 29).

† It was customary at that time to withhold new discoveries and to challenge all comers to solve the problem independently.

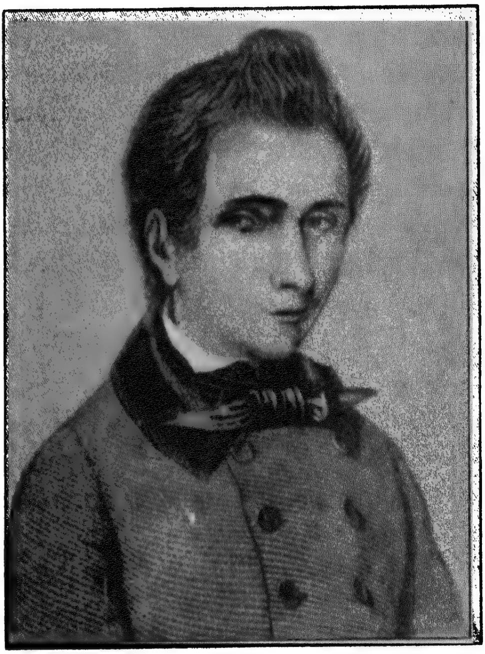
You may be tempted to ask: "How can you boldly assert that it is impossible to find such formulas? Perhaps some day some genius will discover them. All things are possible. Are you sure you don't mean simply that no one has found them yet?" The answer is that we don't merely mean that no one has found them yet; we mean that no one will ever find them because it is impossible for such formulas to exist. If you ask how we can prove something definitely impossible, we reply that you have already done such things yourselves. In section 24, Chapter IV, we showed that it is impossible to find a fraction whose square is exactly 2. We did not say merely that we had examined 10,000 fractions and hadn't yet found one whose square is 2. In fact, we pointed out that such a procedure could not possibly lead to a proof. We actually proved that the existence of a rational number whose square is 2 would lead inevitably to self-contradiction and is, therefore, impossible. All proofs of impossibility are of that general nature. An even simpler example of an impossible task is to find two even numbers whose sum is odd. (See exercise 8, section 24.)



Niels Henrik Abel
1802–1829, Norwegian

Notice that we have not said that the general equation of degree five cannot be solved. In fact, it can be solved by other means, but its roots cannot be given as algebraic expressions in the coefficients. However, the roots of some *particular* equations of degree five or more can be written as algebraic expressions in terms of the coefficients. For example, if in the 5th degree equation above we restrict ourselves to the particular case where $b = c = d = e = 0$, $a \neq 0$, that is, to equations of the form $ax^5 + f = 0$, then we can clearly express one root as

$x = \sqrt[5]{-f/a}$ which is an algebraic expression. Therefore a natural question to raise is: given a definite polynomial equation of degree 5 or more, how can we tell whether or not its roots are expressible as algebraic expressions in its coefficients? * This question was settled by Évariste Galois. The work of Galois was quite original in character and was not well understood at the time because of the sketchy expositions which he presented. Galois' ability was not appreciated by his teachers. In fact, because of various circumstances, including manuscripts lying unread on desks, he received no recognition for his work while he lived. However, what is now called the Galois theory of equations is studied everywhere by advanced students.



Évariste Galois
1811–1832, French

Abel was not yet 27 when he died, leaving behind a wealth of highly original work which stimulated mathematical research for many years after. Galois was killed in 1832 in a duel at the age of less than 21. It is interesting to note that hardening of the arteries is not prerequisite for excellent work in mathematics, although in the tragic case of Galois, rigor mortis was prerequisite for recognition.

39. Simultaneous linear equations. Consider the linear equation $x + y = 7$. It is satisfied by the values $x = 1$, $y = 6$, and by the values $x = 2$, $y = 5$, and by the values $x = 9$, $y = -2$, and so on. Clearly if you choose any value for x the equation is satisfied by that value of x and the value $y = 7 - x$; for example, if we take $x = 15/2$ then the values $x = 15/2$, $y = -1/2$ sat-

* We have here an instance of advanced research arising from exceedingly natural and elementary questions.

isfy the equation. Any pair of values for x and y which satisfies the equation is called a **solution** of the equation. A single linear equation in two variables has infinitely many solutions, one for each value of x. If we consider a system of two equations like

$$x + y = 7 \tag{1}$$

and

$$x - y = 1 \tag{2}$$

we may fairly ask whether these equations have a **common solution,** that is, whether there exists a pair of values for x and y which satisfies both equations simultaneously. The technique of solving such a system of equations is to "eliminate" one of the variables. We reason as follows. If (x,y) is a pair of numbers satisfying both equations (provided there is such a pair), then from (1) we have $y = 7 - x$. Substituting this in (2) we have $x - (7 - x) = 1$, an equation in one variable. Hence, $2x - 7 = 1$ or $x = 4$. But $y = 7 - x = 7 - 4 = 3$. We have thus proved that if (x,y) is a pair of numbers satisfying (1) and (2) then x can only be 4 and y can only be 3. To show, conversely, that this pair of numbers really does satisfy both equations we can merely substitute $x = 4$, $y = 3$ in *both* of the equations and verify it directly.*

It can of course happen that a pair of equations have *no* common solution. For example, consider the system

$$x + y = 5 \tag{3}$$

$$x + y = 6. \tag{4}$$

It is perfectly clear that if x and y are any numbers satisfying (3) they cannot satisfy (4) since their sum cannot be both 5 and 6. Such equations are called **incompatible,** or **inconsistent.**

It can also happen that *every* pair of numbers satisfying one equation will also satisfy the other. For example, consider

$$x + y = 15 \tag{5}$$

$$2x + 2y = 30. \tag{6}$$

Since $2x + 2y = 2(x + y)$ it is clear that if x and y satisfy (5) (that is, their sum is 15), then $2x + 2y = 30$ automatically.

* The student may recall various tricks for performing the elimination of one of the variables, perhaps slightly more rapid than our procedure. But our method is straightforward and can be applied to more general situations as in the next section.

Such a pair of equations may be called **dependent or equivalent.**

A graphical or geometric interpretation of the work of this section and the next will be discussed in Chapter IX, section 84.

EXERCISES

Solve the following systems:

1. $\begin{cases} 2x + 3y = 13 \\ 3x - y = 3. \end{cases}$ **2.** $\begin{cases} 2x + 5y = 5 \\ x - 3y = 8. \end{cases}$ **3.** $\begin{cases} x - y = 1 \\ 2x - 3y = 5. \end{cases}$

4. $\begin{cases} 2x - 3y = 5 \\ 5x + y = 2. \end{cases}$ **5.** $\begin{cases} 3x + 4y = 2 \\ x - 2y = 3. \end{cases}$ **6.** $\begin{cases} 5x - 2y = 3 \\ 2x + 3y = 7. \end{cases}$

7. $\begin{cases} 3x - 2y = 1 \\ 2x + 3y = 3. \end{cases}$ **8.** $\begin{cases} 5x + 2y = 3 \\ 3x - 4y = 7. \end{cases}$ **9.** $\begin{cases} 7x + 3y = 1 \\ 3x - 7y = 7. \end{cases}$

10. $\begin{cases} 5x + 4y = 18 \\ 6x + 2y = 16. \end{cases}$

40. Simultaneous equations of higher degree. We shall take only the case where one equation is linear and the other quadratic. For example, consider the equations

$$x^2 + y^2 = 34 \tag{1}$$

$$x - y = 2. \tag{2}$$

We obtain $x = y + 2$ from (2) and substitute this expression in (1), obtaining

$$\begin{aligned} (y + 2)^2 + y^2 &= 34, \\ y^2 + 4y + 4 + y^2 &= 34, \\ 2y^2 + 4y - 30 &= 0, \\ y^2 + 2y - 15 &= 0, \\ (y + 5)(y - 3) &= 0, \\ y = -5 \text{ or } y &= 3. \end{aligned}$$

For $y = -5$ we obtain from (2) the value $x = -3$ and for $y = 3$ we obtain the value $x = 5$. To see whether both the pair $(x = -3, y = -5)$ and the pair $(x = 5, y = 3)$ satisfy the system we substitute in (1) and (2).

The problem of solving simultaneously two polynomials in x and y of any degree is in general quite difficult. The procedure used above and in the preceding section was to express one variable in terms of the other from one equation and substitute in the other equation, thus eliminating one of the variables and obtaining an equation in the other variable alone which could then be solved. But this was easy because we had, in the above cases, a

linear equation from which we could obtain an expression for y in terms of x, or x in terms of y. From equations like $5x^3 + x^2y + 7xy^2 + 11x + 13y^3 = 17$ and $x^5 + 3xy^3 - y^4 = 19$, it would not be so easy to eliminate x or y. The mathematician's tendency to seek general theorems and methods would inevitably lead him to attack the general problem of solving simultaneously any number of equations of any degree in any number of unknowns. This problem leads to interesting, modern, and difficult researches in higher algebra, called elimination theory. Here again we have an example of advanced mathematics which springs from natural and elementary questions.

EXERCISES

Solve the following systems:

1. $\begin{cases} x^2 + y^2 = 25 \\ 2x + y = 10. \end{cases}$

2. $\begin{cases} x^2 + y^2 = 1 + 4x \\ 2x - y = 4. \end{cases}$

3. $\begin{cases} x^2 + y^2 = 13 \\ 2x - y = 4. \end{cases}$

4. $\begin{cases} xy = 1 \\ 3y - 5x = 2. \end{cases}$

5. $\begin{cases} 7y^2 - 6xy = 8 \\ 2y - 3x = 5. \end{cases}$

6. $\begin{cases} y^2 = 4x \\ y + 2x = 4. \end{cases}$

7. $\begin{cases} x + y = 7 \\ xy = 12. \end{cases}$

8. $\begin{cases} y^2 = 4x \\ x + y = 3. \end{cases}$

9. $\begin{cases} x^2 + y^2 = 25 \\ 4x - 3y = 0. \end{cases}$

10. $\begin{cases} x^2 + y^2 = 41 \\ x + y = 9. \end{cases}$

11. $\begin{cases} x^2 - y^2 = 5 \\ 3x + y = 11. \end{cases}$

12. $\begin{cases} xy = 2 \\ x^2 = y. \end{cases}$

41. Problems leading to the solution of equations. The student is already familiar with verbal problems which lead to the solution of equations. One must read the problem carefully, choose symbols to represent the various quantities involved, and use the given relations among these quantities to set up the equation or equations. The problem is essentially nothing but that of translating from the clumsy language of everyday prose into the much more convenient language of algebra. Hence the student should carefully write down a systematic list or vocabulary of all symbols to be used and express all the quantities to be considered in terms of these symbols.* If you fail to write down what your symbols stand for, you are writing in a secret code.

* Various schemes of arrangement of this vocabulary on the page, such as "boxes" or other bookkeeping devices seem to be popular. They may save a few seconds but are inessential and often serve only to obscure the reasoning.

Example 1. Within a flower garden 6 yards wide and 12 yards long, we want to pave a path of uniform width around the boundary so as to leave an area of 40 square yards for flowers. How wide shall we make the path?

Let x = width of path, measured in yards. Then the width of the actual flower plot is $6 - 2x$, and its length is $12 - 2x$. The area of this flower plot is therefore $(6 - 2x)(12 - 2x)$. But this area is to be 40 square yards. Hence

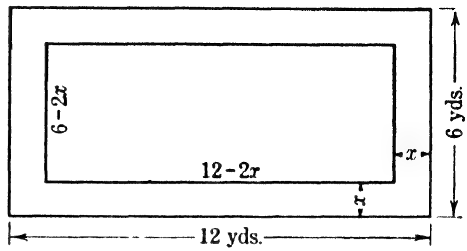


FIG. 28

(1) $$(6 - 2x)(12 - 2x) = 40.$$

Solving this equation we have

$$72 - 36x + 4x^2 = 40,$$

or, $$x^2 - 9x + 8 = 0$$

or, $$(x - 1)(x - 8) = 0.$$

Hence, $$x = 1 \text{ or } x = 8.$$

Clearly $x = 1$ satisfies the conditions of the problem for the flower plot then has the dimensions $6 - 2 = 4$ and $12 - 2 = 10$ which yield an area of 40. But the root $x = 8$ clearly does not satisfy the conditions of the problem since you cannot pave a path 8 yards wide in a garden only 6 yards wide. However, 8 is a bona fide root of equation (1). How can this happen? The answer is simply that in setting up the equation (1) we reasoned that *if* x satisfies the conditions of the verbal problem, *then* x will have to satisfy the equation (1). But the *converse* of this proposition need not be so. Thus it should not surprise us if there is a value of x which satisfies the equation without satisfying the problem. In all verbal problems it is therefore essential to check the roots of the equation against the actual verbal problem before accepting them.

Example 2. It is found experimentally that if a corpse is dropped from the top of a vertical cliff, it falls $16t^2$ feet in t seconds. (This is known as the falling body problem.) If the cliff is 144 ft. high, in how many seconds will the body reach the bottom?

Clearly $16\,t^2 = 144$ or $t^2 = 9$. Hence $t = \pm 3$. But -3 seconds has no meaning for our problem. Thus 3 seconds is the answer. However, -3 is a root of our equation. Explain this phenomenon.

Example 3. The following was proposed to Leonardo of Pisa, a famous mathematician, about 1200 A.D., as a difficult problem.*

If A gets from B seven denare,† then A's sum is five-fold B's; if B gets from A five denare, then B's sum is seven-fold A's. How much has each?

Let x represent A's original sum and let y represent B's original sum. Then clearly,

$$x + 7 = 5(y - 7)$$

and

$$7(x - 5) = y + 5.$$

Solving these equations simultaneously we obtain $x = 7\frac{2}{17}$ denare and $y = 9\frac{14}{17}$ denare.

The fact that this problem was considered difficult indicates the importance of an adequate symbolism. The reader has only to try to solve this problem, or some other problem, without using any symbolic notation to be convinced of this. No convenient algebraic notation was used about 1200 A.D.

EXERCISES

1. The area of a rectangular garden is 48 square yards. The perimeter is 28 yards. Find the length and width of the field.

2. A stone is dropped from a vertical cliff 96 ft. high. Using the formula

$$s = 16\,t^2 \tag{1}$$

for the distance s, measured in feet, through which the body will fall in t seconds, find the number of seconds it takes for the stone to reach bottom.

3. A can do a certain job in 6 days and B can do it in 10 days. How long will it take them to do it together?

4. A man is 24 years older than his son. Eight years ago he was twice as old as his son was. What are their present ages?

5. A solution of acid is 75% pure. How many grams of acid must be added to 48 grams of this solution in order to make the resulting solution 76% pure?

6. A stone is dropped from a cliff overlooking a lake, and the sound of the impact on the water is heard 9 seconds after the stone was dropped. If the velocity

* Cajori, *History of Mathematics*, 2nd Edition, p. 123.

† An ancient coin.

of sound is 1024 ft. per sec. and a body falls 16 t^2 feet in t seconds, how high is the cliff above the lake?

7. A man invests part of \$5000 at 4% interest and the rest at 6%. His total annual income from both investments is \$280. How much has he invested at each rate?

8. A boat, operating uniformly at full power, goes 5 miles downstream in 60 minutes and returns in 90 minutes. What would the speed of the boat be in still water and what is the rate of the current?

9. A can do a job in 8 days, and A and B together do it in 4 days. How long would it take B to do it alone?

10. A collection of nickels and quarters contains 30 coins. Their total value is \$3.70. How many of each kind of coin are there?

11. A rectangular field is four times as long as it is wide. If it were 5 feet shorter and 2 feet wider its area would be increased by 20 square feet. Find its length and width.

12. An automobile travels 120 miles. A second automobile travels 10 miles per hour faster than the first and makes the same trip in 2 hours less time. Find the speed of each.

13. A and B do a job together in 6 days. A alone can do it in 5 days less than it takes B to do it alone. How long would it take each to do the job alone?

14. The two legs of a right triangle differ by 7 feet. The hypotenuse is 13 feet long. Find the legs.

15. The area of a rectangle is 400 square feet. The perimeter is 100 ft. Find the length and width.

16. A man's income is \$25,000. A certain percentage is taken off for State taxes and the same percentage is taken off what is left for Federal taxes. After both taxes are paid, the man has \$24,010. What is the tax rate?

17. A freight train, running at the rate of 20 miles per hour leaves a station three hours before an express which travels in the same direction at the rate of 50 miles per hour. How long after the express leaves will it catch the freight and how far from the station will they be?

18. Achilles races with a Tortoise, giving the Tortoise a handicap of 990 yards. Achilles runs at the rate of 500 yards per minute while the Tortoise runs at the rate of 5 yards per minute. How long will it take Achilles to catch the Tortoise?

19. How many gallons of cotton seed oil must be added to 45 gallons of a solution of olive oil which is 90% olive oil in order to make the resulting solution 80% olive oil?

20. The age of Diophantus, a brilliant Greek mathematician of about 250 A.D., may be calculated from an epitaph which runs as follows: Diophantus passed one sixth of his life in childhood, one twelfth in youth, and one seventh more as a bachelor; five years after his marriage was born a son who died four years before his father, at half his father's age.

REFERENCES

E. T. Bell, *Men of Mathematics*, Simon and Schuster, N. Y., 1937. (Short biographies of Galois, Abel, Gauss, and many others.)

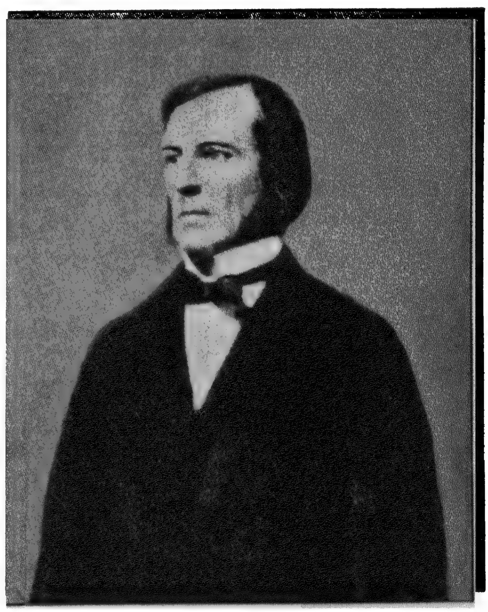
H. B. Fine, *College Algebra*, Ginn, Boston, 1904. (Among other things, explanations of the processes of long division, and the extraction of square roots of polynomials.)

Histories of Mathematics, as before.

Chapter VI

THE ALGEBRA OF LOGIC AND OTHER ALGEBRAS

42. Introduction. In the last three chapters we have been engaged in showing: (1) that algebra has the same kind of logical structure as any other abstract mathematical science; that is, it starts with undefined terms and unproved assumptions or postulates and then proceeds by defining new terms and proving new statements; (2) that the algebraic manipulations with which you have been familiar for years in a mechanical way are logical consequences of our postulates and definitions; (3) that our choice of postulates and definitions has been guided by the concrete applications that we had in mind for our logical system; that is, our postulates and definitions have been suggested by experience. Briefly we have been concerned with what may be called the *logic of algebra*, where by algebra we mean the familiar algebra of numbers.



George Boole

1815–1864, English

We have pointed out repeatedly (see sections 16 and 17) that we might very well have used different postulates and different definitions, especially if we had in mind different sorts of applications. We shall now take up briefly actual examples of algebras which in some respects resemble the algebra of numbers and in other respects differ from it sharply, and which really have important applications.

43. The algebra of sets or the algebra of logic. The algebra we shall discuss in this section is variously called the **algebra of sets,** the **algebra of classes,** the **algebra of logic,** or **Boolean algebra,** because it was developed by G. Boole about 1850.

Let us begin by considering any definite *collection* or *set* or *class* of objects. For example,

(1) a class of four people: Smith, Jones, Brown, and Robinson;

or,

(2) the set of all people in New York City;

or,

(3) the set of all points on a given page.

*Whatever set we choose to consider, let that be the basic set from here on.** The objects in this basic collection will be called **elements.** By a **set** we will mean some class of *these* elements chosen from the basic collection. Thus if we choose the first example above as our basic set, a class consisting of Smith and Robinson is a set; the class consisting of Jones and Smith is a different set. If we choose the second example above as our basic set, the set of people in New York 21 years of age is a set; the set of people in New York with blue eyes is a different set. If we take the third example as our basic set, the class of points within the curve in Fig. 29 is a set; the class of points outside the curve is a different set.

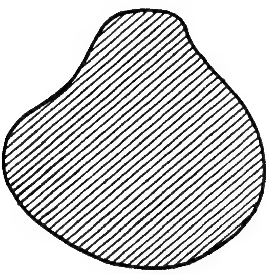


Fig. 29

It will be convenient to regard the entire basic collection as a set and to denote it by the symbol 1. This symbol is not to be confused with the number one—it simply means the entire basic collection of objects we are considering. It will also be convenient to consider the set which has no elements in it. For example, the set of people in New York who are more than 30 ft. tall is a set with no elements. *This set is called the* ***empty set*** *and is denoted by the symbol* 0. This symbol is not to be confused with the number 0. We conceive of just one empty set; it is a matter of indifference to us whether a set has no people in it or no dogs.

* The diagrams will be most easily interpreted as referring to the basic collection (3) of all points on a given page.

Two sets a and b are called **equal** if and only if they have exactly the same elements or members; that is, if every element of a is an element of b and every element of b is an element of a.

By the **sum** of two sets we shall mean simply the set composed of all the elements belonging to one or the other or both of these two sets. For example, the sum of the set consisting of Smith and Robinson and the set consisting of Jones and Smith will be the set consisting of Smith, Robinson, and Jones. If we denote sets by letters $a, b, \ldots$, we shall denote the sum of two sets a and b by $a + b$. If a and b are the sets of points on the page (Fig. 30) which are shaded horizontally and vertically respectively, then the set $a + b$ is the whole set of points shaded in any way.

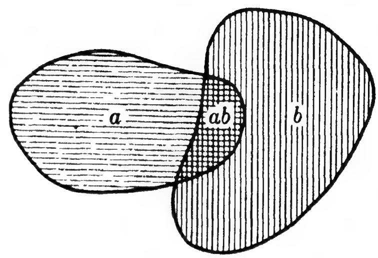


Fig. 30

By the **product** of two sets we shall mean the set composed of those elements which belong simultaneously to both of the given sets—that is, their common part. For example, the product of the set consisting of Smith and Jones and the set consisting of Smith and Robinson is the set consisting of Smith alone. If a and b are the horizontally and vertically shaded sets of points in Fig. 30, respectively, the product denoted by ab, or $a \cdot b$, will be the set which is shaded crosswise.

For example, suppose in our first illustrative basic set we denote by a the set consisting of Smith and Jones and by b the set consisting of Brown and Robinson. Then we may write $a + b = 1$ because the sum of these two sets is the entire basic set. We may also write $ab = 0$ because these two sets have no elements in common; that is, their product or common part is the empty set.

If every element of one set a is also an element of another set b, then a is said to **be contained within** b; we write $a \subset b$, or $b \supset a$ (read "a is contained in b" or "b contains a"). For example, the set consisting of Smith and Jones is contained within the set consisting of Smith, Jones, and Brown. Note that the relation $\subset$ which holds between sets is entirely different from the relation $<$ which holds between numbers, although they have some similar properties (see, for example, property X below). Be careful not to confuse these two relations.

We can also introduce the idea of **subtraction.*** Let $a - b$ denote the set of those elements of a which are not in b, (Fig. 31). For example, the set of *all* elements not in the set x is denoted by $1 - x$; that is, $1 - x$ represents the set of all elements in the entire basic collection 1 which are not in x.

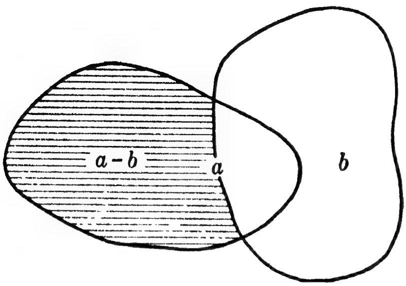


Fig. 31

We can now verify that sets have the following properties, all letters representing sets in a given basic collection.

I. To each pair of sets a and b, in that order, there is a third set denoted by $a + b$, called the sum of a and b.

II. $a + b = b + a$.

III. $(a + b) + c = a + (b + c)$.

IV. To every pair of sets a and b, in that order, there is a third set denoted by ab or $a \cdot b$, called the product of a and b.

V. $ab = ba$.

VI. $a(bc) = (ab)c$.

VII. $a(b + c) = ab + ac$.

These can be verified easily by examining Fig. 32 carefully and keeping firmly in mind the meaning of sum and product. Shading with colored crayons will help you in verifying these facts. For example, VII asserts that if you first take the sum of the sets

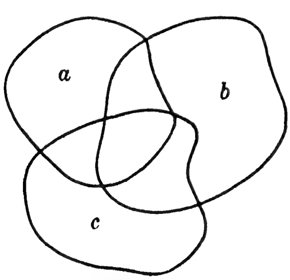


Fig. 32

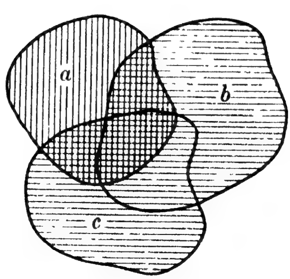


Fig. 33

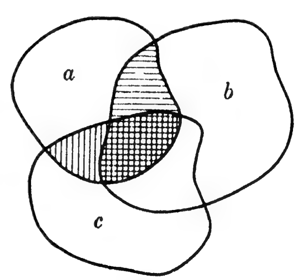


Fig. 34

* The usual definition of subtraction, namely "$a - b$ means the set x such that $b + x = a$" would not be suitable here. For such a set would not exist unless $b \subset a$ and even then would, in general, not be unique. Similarly we do not find it suitable to introduce the definition of division "$a \div b$ means the set x such that $bx = a$." For such a set would not exist unless $b \supset a$ and even then would not be unique. The student should verify these statements by experimenting with diagrams.

b and c (shaded horizontally in Fig. 33) and consider the common part of this set $b + c$ with a (shaded vertically in Fig. 33) then you get exactly the same resulting set (shaded crosswise in Fig. 33) as if you had first taken the set common to a and b (shaded horizontally in Fig. 34) and the set common to a and c (shaded vertically in Fig. 34) and then taken their sum $ab + ac$ (the entire shaded set in Fig. 34).

We also see easily that the following properties hold for sets.

VIII. There is a set 0 (the empty set) such that

$$(a)\quad 0 + a = a$$
$$(b)\quad 0 \cdot a = 0$$

for any set a.

IX. There is a set 1 (the entire basic set), such that

$$1 \cdot a = a$$

for any set a.

X. If $a \subset b$ and $b \subset c$ then $a \subset c$. (See Fig. 35.)

If the reader will now compare the properties of this algebra of sets with the correspondingly numbered properties of the algebra of numbers given in section 28, he will see that they are quite the same so far (with $\subset$ taking the place of $<$ and the word "set" taking the place of the word "number") except for VIIIc in section 28. The property corresponding to VIIIc for sets is not true. For we can have sets $a \neq 0$ (that is, not empty) and $b \neq 0$ such that $ab = 0$ (see Fig. 36);

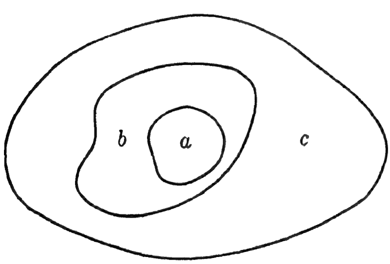


FIG. 35

that is to say, they have no elements in common, or their common part is the empty set. Some students are of the opinion that if a is the set of mathematics instructors and b is the set of sane and kind-hearted people then $ab = 0$. The algebra of sets has other peculiar properties not possessed by the algebra of numbers. For example, for any set a we have $a \subset 1$; while the corresponding statement for numbers ("for any number a, we

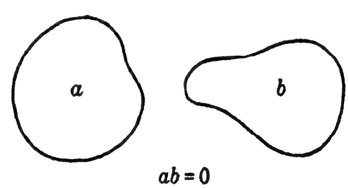


FIG. 36

have $a < 1$") is clearly false. We also have peculiar formulas like the following which hold for all sets, but certainly not for all numbers.

XI. $a + a = a$. In particular, $1 + 1 = 1$.

XII. $aa = a$

XIII. $ab \subset a$.*

XIV. If $x \subset y$, then $y + x = y$. In particular, since $a \subset 1$ for all sets a, we have $1 + a = 1$.

XV. For every set x there is a corresponding set x' such that $x + x' = 1$ and $xx' = 0$. (This x' is the set $1 - x$.)

Remark. If we interchange $+$ and $\cdot$, 0 and 1, $\subset$ and $\supset$, in any correct formula, we obtain another correct formula in the algebra of sets. For example, interchanging $+$ and $\cdot$ in VII we obtain $a + (b \cdot c) = (a + b) \cdot (a + c)$ which is true for sets (verify this on a figure like Fig. 32) but is obviously false for numbers.

The algebra of sets therefore cannot be the same as the algebra of numbers. However, it is worth noting that whatever theorems can be deduced by using *only* those properties which both have in common will be true in either system.

We might construct an abstract mathematical science with "set," "sum," "product," "contained in," "equals," as undefined terms and the above properties as postulates. Actual sets of objects would provide a concrete interpretation of this "algebra." What applications can this queer algebra have? It is clear that we might be able to use it in any situation in which we are concerned with collections of objects whatever the nature of these objects may be. The uses of this algebra in higher mathematics are extremely numerous. But the following elementary application explains why it is sometimes called the *Algebra of Logic.*

The reader may have already observed a resemblance between the diagrams of sets of points on the page which have been drawn here and the diagrams used to test reasoning in section 4. Let us see how the familiar statements of logic can be symbolized in our new algebra.

Let a be the set of all freshmen and b the set of all intelligent

* In order to reconcile XIII with the statement that ab may be 0 even though $a \neq 0$ and $b \neq 0$ we make the agreement that the empty set shall be regarded as contained in every set.

people. Then the statement "all freshmen are intelligent" may be written (see Fig. 37) as

$$ab = a,$$

or $a \subset b$, or $a - b = 0$. "Some freshmen are intelligent" may be written

$$ab \neq 0.$$

"No freshmen are intelligent" becomes

$$ab = 0$$

or $a \subset 1 - b$. "Some freshmen are not intelligent" becomes

$$a(1 - b) \neq 0.$$

It is possible to express the arguments of traditional logic in terms of this algebra, and one can actually reason by manipulating the symbols according to the rules (postulates) of this alge-

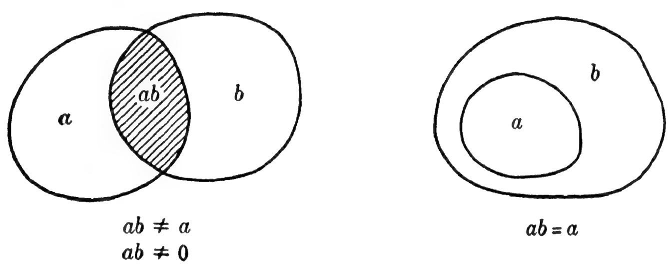


FIG. 37

bra. It is clear, for example, that the reasoning of example 1, section 4, is merely a direct application of property X of the algebra of sets. In effect, traditional logic can be reduced to the study of this "algebra of logic," or other symbolic "algebras" like it.

Example. Consider the following valid argument:

Hypothesis: (1) All freshmen are undergraduates;
(2) No undergraduates are intelligent;

Conclusion: Therefore no freshmen are intelligent.

The conclusion of this argument can be obtained algebraically as follows. Let a, b, c be the classes of freshmen, undergraduates, and intelligent people, respectively. Then the two parts of the hypothesis are written as

(1) $$ab = a$$

and

(2) $$bc = 0$$

respectively. We obtain the conclusion $ac = 0$ as follows. By (1),

$$\begin{aligned} ac &= (ab)c && \text{(substituting } ab \text{ for } a \text{ which is its equal by hypothesis (1))} \\ &= a(bc) && \text{(by the associative law VI)} \\ &= a \cdot 0 && \text{(by hypothesis (2))} \\ &= 0 && \text{(by VIII}b\text{)} \end{aligned}$$

which was to be proved.

It is worth noting that this "symbolic logic" grew up not as a mere plaything of the mathematician but as a response to a real need. Many paradoxes and confusions in logic were found to be due to the deficiencies of ordinary language.* Ordinary language evolves from the everyday needs of people to express their ideas, and this does not usually require much in the way of precision or subtle distinctions. Therefore it is natural to expect ordinary language to be poorly adapted to the needs of careful logic. Thus a symbolism was created for the purpose. We have already remarked that the development of satisfactory symbolisms is not without importance. We shall see some aspects of this in the next chapter.

In deducing further results from postulates I–XV you will have to consult the postulates carefully in order to be sure that your steps are justified, since you cannot rely on your familiarity with algebraic manipulations here as you may have done in the case of the algebra of numbers.

EXERCISES

1. By shading the indicated sets suitably, on diagrams like that of Fig. 32, verify: (*a*) property III; (*b*) property VI; (*c*) property VII; (*d*) the formula $a + (bc) = (a + b) \cdot (a + c)$.

2. Defining $a \subset b$ to mean $ab = a$, and so on, prove that if $a \subset b$ and $b \subset c$ then $a \subset c$. (Hint: use the associative law VI and substitution. Imitate the example in the text, as far as possible.) This shows that the property X need not have been assumed since we can prove it on the basis of the other postulates.

* A trivial example of verbal confusion is found in the well-known example: "no cat has 8 tails; every cat has one more tail than no cat; therefore every cat has 9 tails."

3. Show that if $ab = a$ and $ac \neq 0$, then $bc \neq 0$. (Hint: show that the set $a(bc) \neq 0$. From this the conclusion follows. Why?)

4. Express each statement in algebraic terms and obtain algebraically the conclusion of the argument given in:

(*a*) section 4, example 3.
(*b*) section 4, exercise 4.
(*c*) section 4, exercise 6.
(*d*) section 4, exercise 7.
(*e*) section 4, exercise 12(i).

5. From the properties I–XV of sets listed above, deduce the following theorems, all letters representing arbitrary sets:

(*a*) $a + ab = a$. (Hint: use XIII and XIV.)
(*b*) $a(a + b) = a$. (Hint: use (*a*).)
(*c*) $(a + b)(a + c) = a + (bc)$. (Hint: use (*a*) and (*b*).)

44. Algebra on a dial. In the preceding section we discussed an example of an algebra which resembled the algebra of numbers in some respects but differed from it in others. In this section we shall discuss still other algebras different from the usual algebra of numbers.

While you may have spent a large part of your academic career watching a clock, you may never have noticed that a curious algebra can be constructed from the clockface, or, for that matter, any dial.

Algebra of integers modulo five. Consider a dial like that of Fig. 38. Let the **elements** or **numbers** of our algebra be the five numbers 0, 1, 2, 3, 4 *alone*. The "sum" of any two of these "numbers" will be interpreted as follows. Choosing a positive direction of rotation, as in Fig. 38, $2 + 1$ shall be obtained by starting at 0, rotating positively 2 spaces, and then rotating positively 1 space, arriving at 3; hence $2 + 1 = 3$. That is the **sum** of any two of our five numbers shall mean the number to which the dial hand points after performing the two indicated rotations in succession, starting from 0. This is analogous to addition of ordinary numbers, except that we are dealing with a circle instead of a straight line. Zero is interpreted to mean "rotate not at all." The following table of addition is easily verified:

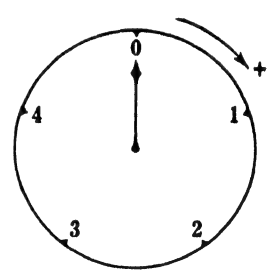


FIG. 38

$0 + 0 = 0$	$1 + 0 = 1$	$2 + 0 = 2$	$3 + 0 = 3$	$4 + 0 = 4$
$0 + 1 = 1$	$1 + 1 = 2$	$2 + 1 = 3$	$3 + 1 = 4$	$4 + 1 = 0$
$0 + 2 = 2$	$1 + 2 = 3$	$2 + 2 = 4$	$3 + 2 = 0$	$4 + 2 = 1$
$0 + 3 = 3$	$1 + 3 = 4$	$2 + 3 = 0$	$3 + 3 = 1$	$4 + 3 = 2$
$0 + 4 = 4$	$1 + 4 = 0$	$2 + 4 = 1$	$3 + 4 = 2$	$4 + 4 = 3.$

The commutative and associative laws for addition may now be proved by merely verifying all possible combinations, since there is only a finite number of "numbers." Thus $3 + 4 = 2$ and $4 + 3 = 2$, for example; this verifies the commutative law for one instance. Similarly $4 + (3 + 1) = 4 + 4 = 3$ while $(4 + 3) + 1 = 2 + 1 = 3$; this verifies the associative law for one instance. The results of our table of addition appear insane only if you persist mistakenly in interpreting our symbols in the usual way instead of the way in which they are intended to be interpreted.

Multiplication may be defined as in section 13. Thus $4 \cdot 3 = 3 + 3 + 3 + 3 = 2$, and $3 \cdot 4 = 4 + 4 + 4 = 2$, as may be seen from the table of "addition." Subtraction and division may be be defined as usual. Thus $2 - 4$ means a number x such that $4 + x = 2$; from the addition table we see that $x = 3$ since $4 + 3 = 2$. Hence $2 - 4 = 3$. Similarly $2 \div 4$ means a number x such that $4x = 2$. Thus $2 \div 4 = 3$ since $4 \cdot 3 = 2$.

This algebra is called the *algebra of integers modulo 5* because there are only five "numbers" in it, namely 0, 1, 2, 3, 4. It has many of the properties of ordinary algebra such as the associative, commutative, and distributive laws, but clearly differs from ordinary algebra in some respects.

EXERCISES

1. Construct a table of multiplication for our 5 numbers.

2. Verify from the table in exercise 1 that the commutative and associative laws for multiplication hold for four instances each.

3. Verify the distributive law for four instances.

Calculate in the algebra of integers modulo 5:

4. $1 \div 3$. **5.** $1 \div 4$. **6.** $1 \div 2$. **7.** $3 \div 4$. **8.** $4 \div 3$.

9. $3 - 4$. **10.** $2 - 3$. **11.** $1 - 4$.

Algebra of integers modulo six. Clearly, an algebra can be constructed similarly from a dial with any number of "numbers" on

it. Consider the algebra on a dial with six numbers (Fig. 39), that is, *the algebra of integers modulo six.* Then $1 + 1 = 2$, $4 + 4 = 2$, $3 + 3 = 0$, and so on. Hence $2 \cdot 1 = 2$, $2 \cdot 4 = 2$, $2 \cdot 3 = 0$. Notice that $2 \neq 0$ and $3 \neq 0$, but $2 \cdot 3 = 0$. This could happen neither in the ordinary algebra of numbers nor in the algebra of integers modulo 5 (Fig. 38). As a consequence of this phenomenon we shall find the concept of division very troublesome in this algebra. For example, $1 \div 2$ does not exist at all, while $2 \div 2$ may be either 1 or 4, since $2 \cdot 4 = 2$ and $2 \cdot 1 = 2$. Thus some quotients do not exist while others are not uniquely determined.

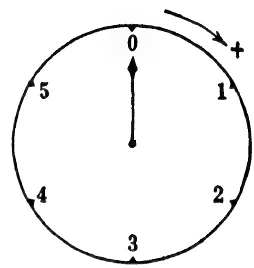


Fig. 39

Remark. A natural number greater than 1 is called a **prime** if it has no factors other than itself and one. The first few prime numbers are 2, 3, 5, 7, 11, 13, $\cdots$. The troubles with division just described can occur in the algebra on a dial only if the number of numbers on the dial is not prime; we shall not prove this here.

EXERCISES

12. Construct tables of addition and multiplication for the algebra of Fig. 39.

13. By means of these tables, verify the commutative and associative laws for addition and multiplication, for four instances each.

14. Verify the distributive law for four instances.

15. Show that the following quotients do not exist:

(*a*) $1 \div 2$. (*b*) $3 \div 2$. (*c*) $4 \div 3$.

16. Show that $2 \div 2 = 1$ and $2 \div 2 = 4$.

17. Show that $4 \div 2 = 2$ and $4 \div 2 = 5$.

Calculate in the algebra of integers modulo 6:

18. $4 \div 5$. **19.** $2 \div 4$. **20.** $1 \div 3$. **21.** $3 \div 3$.

22. $2 - 5$. **23.** $3 - 5$. **24.** $3 - 4$.

Conclusion. In this section and the last we have brought out again that we might well develop algebras different from the usual algebra of numbers if we have different applications in mind. For example, an algebra constructed so as to apply to

a clock-dial would be quite different from our usual algebra which is constructed for the sake of other applications.

REFERENCES

E. T. Bell, *Men of Mathematics*, Simon and Schuster, N. Y., 1937. (Biography of Boole.)

A. A. Bennett & C. A. Baylis, *Formal Logic*, Prentice Hall, N. Y., 1939. (Especially Chapter VII.)

M. R. Cohen & E. Nagel, *Introduction to Logic and Scientific Method*, Harcourt Brace, N. Y., 1934. (Chapter VI.)

Lewis and Langford, *Symbolic Logic*, The Century Co., New York, London, 1932. (Chapter II.)

Histories of Mathematics, as before.

Chapter VII

ARITHMETIC, EXPONENTS, AND LOGARITHMS

45. Introduction. Algebra, as you know it, dates roughly from the 13th to the 17th centuries, during which time algebraic symbolism passed through many stages of development, although some of it goes back thousands of years. The first symbolic advances over ordinary prose were mere abbreviations. For example, in his *Ars Magna de Rebus Algebraicis* (1545), Cardan wrote "cubus p 6. rebus aequalis 20" for $x^3 + 6x = 20$. The equal sign $=$ is found for the first time in print in *The Whetstone of Witte* by Robert Recorde, published in 1557. The symbol ℞ ("radix," root) was often used; thus, ℞ cu. 8 would be written for $\sqrt[3]{8}$. In the early 17th century we still find *aaaaa* written for a^5. The ancients had no adequate arithmetical or algebraic symbolism and it is probable that this greatly handicapped their development of arithmetical and algebraic ideas.

Systems of notation for numbers seem to have been almost universally influenced by the anatomical accident that most of us possess two hands with five fingers on each. There is scarcely any doubt that the human race learned to count on its fingers, just as a child does. In fact, various elaborate schemes of finger-reckoning are still in use today among unlettered people. Thus, the numbers five and (especially) ten have always been important. Some tribes, the Mayas of Central America, for example, gave twenty a special place and used a word for twenty meaning "a whole person," thus indicating that they counted on both fingers and toes. Other indications that twenty was once considered important are found in the English word "score," and in the French "quatre-vingt" (four twenties) for eighty. But ten is more commonly regarded as basic. The Egyptians, for example, used the symbol ∩ for ten and wrote twenty-three as ∩∩|||. Similarly, the Romans had special symbols for five and ten. For example, they wrote eighteen as XVIII. The no-

tation you learned in school is called the Arabic system, (although it was probably developed by the Hindus about 500 A.D.), because it was brought into Europe by the Arabs after they overran Spain in the 8th century. The earliest Arabic manuscripts using this system date from the 9th century. However, it was not widely used in Europe until the 15th century. The Arabic or Hindu-Arabic notation also shows the influence of the number 10 but in an essentially different way. Thus 7346 means $7 \cdot 10^3 + 3 \cdot 10^2 + 4 \cdot 10 + 6$.

If our sole concern with numbers were to record them, as, say, historians record dates, then one system of notation would be pretty much as good as another. But if we are going to add, multiply and so on, then some systems are clearly superior to others. If anyone doubts it, let him multiply 1748 by 34 as he was taught in school and then try to multiply the same numbers MDCCXLVIII and XXXIV using only the Roman numerals. It is our simple arabic notation which makes it possible for a child of ten to do arithmetical problems which would have taxed the ability of a Roman senator. The ancients, of course, used the abacus, counting boards with pebbles, and other devices. In fact, the word "calculate" comes from the Latin word "calculus" meaning "pebble." Despite its obvious advantages over previous methods, the arabic system was adopted generally in Europe only after a long struggle against the repressive force of tradition and resistance to change. In fact, the new notation was adopted by the enlightened merchants long before the "learned" authorities became reconciled to it. It made it possible for the general public, instead of experts only, to learn arithmetic. Its importance in the operation of our complicated modern industrial and commercial civilization is easy to see.

46. Operations with arabic numerals. The convenience of the arabic system is due to the fact that it is a cleverly devised *positional* notation. That is, we need only the ten symbols or digits (notice the word "digit," meaning finger) 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, other numbers being written as combinations of these in which the position of each digit tells us how it is to be interpreted; thus

$$345 = 3 \cdot 10^2 + 4 \cdot 10 + 5$$

while

$$534 = 5 \cdot 10^2 + 3 \cdot 10 + 4.$$

A number expressed in positional notation is essentially a sum of terms, each term being a power of ten multiplied by a coefficient which is one of the digits from 0 to 9 inclusive. The powers of ten occur in descending order. In school, we spoke of the "units place," "tens place," "hundreds place," and so on.

For example, suppose we add 378 and 145 as we learned to do it in school:

$$\begin{array}{r} {\scriptstyle 1\,1\ } \\ 378 \\ \underline{145} \\ 523. \end{array}$$

Let us examine more closely the reasons for the correctness of this scheme and how the trick of "carrying" arises. This will become apparent if we express these numbers in powers of ten and add them by using only the familiar postulates of algebra. We want the sum

$$(3 \cdot 10^2 + 7 \cdot 10 + 8) + (1 \cdot 10^2 + 4 \cdot 10 + 5).$$

By the generalized associative law we can remove parentheses and by the commutative and generalized associative laws for addition we obtain

$$(3 \cdot 10^2 + 1 \cdot 10^2) + (7 \cdot 10 + 4 \cdot 10) + (8 + 5).$$

The generalized distributive law then allows us to write

$$(3 + 1)10^2 + (7 + 4)10 + (8 + 5)$$

or

$$4 \cdot 10^2 + 11 \cdot 10 + 13.$$

This is not yet in positional notation since the coefficients of some of the terms are greater than 9. Hence we write,

$$4 \cdot 10^2 + (10 + 1)10 + (10 + 3),$$

which by the generalized distributive and associative laws, becomes

$$(4 \cdot 10^2 + 10^2) + (10 + 10) + 3.$$

Using the generalized distributive law again we have

$$(4 + 1)10^2 + (1 + 1)10 + 3 = 5 \cdot 10^2 + 2 \cdot 10 + 3 = 523.$$

Exercise. Find the steps above in which "carrying" took place.

In the same way let us examine the process of multiplication.

Consider the following example done by the device you learned by rote as children:

$$\begin{array}{r} 45 \\ \underline{24} \\ 180 \\ \underline{90} \\ 1080. \end{array}$$

Now let us derive this result from the postulates. We want the product

$$(4 \cdot 10 + 5)(2 \cdot 10 + 4).$$

Regarding the first parenthesis as a single quantity for the moment, the distributive law allows us to write

$$(4 \cdot 10 + 5)2 \cdot 10 + (4 \cdot 10 + 5)4.$$

Using the generalized distributive law we have

$$4 \cdot 10 \cdot 2 \cdot 10 + 5 \cdot 2 \cdot 10 + 4 \cdot 10 \cdot 4 + 5 \cdot 4.$$

By the commutative law for multiplication, we have

$$\begin{aligned} &4 \cdot 2 \cdot 10 \cdot 10 + 5 \cdot 2 \cdot 10 + 4 \cdot 4 \cdot 10 + 5 \cdot 4 \\ &= 8 \cdot 10^2 + 10^2 + 16 \cdot 10 + 2 \cdot 10 && \text{(Reason?)} \\ &= (8 + 1) \cdot 10^2 + (16 + 2) \cdot 10 && \text{(Reason?)} \\ &= 9 \cdot 10^2 + (10 + 8) \cdot 10 && \text{(Reason?)} \\ &= 9 \cdot 10^2 + 1 \cdot 10^2 + 8 \cdot 10 && \text{(Reason?)} \\ &= (9 + 1) \cdot 10^2 + 8 \cdot 10 && \text{(Reason?)} \\ &= 1 \cdot 10^3 + 8 \cdot 10 = 1 \cdot 10^3 + 0 \cdot 10^2 + 8 \cdot 10 + 0 = 1080. \end{aligned}$$

Notice the importance of the rather sophisticated invention of a symbol for zero in positional notation. Without it, we would have difficulty in distinguishing between 1080, 108, 180, 1800, and 18. On an abacus (see Fig. 13) an empty line plays the part of zero. In fact, the abacus contains the essence of positional notation, since one line is used for units, the next for tens, and so on, and in using the abacus one "carries" from one line to the next.

The processes of division and the extraction of square roots might be similarly analyzed but as they are somewhat more complicated we shall not do so here.

The resemblance between the mechanical processes for operations with polynomials and for operations with arabic numerals may be traced to the fact that the number 378 is nothing more

than a particular value of the polynomial $3x^2 + 7x + 8$; namely the value obtained when $x = 10$. The only change is in the carrying process which takes place in the case of numbers. Thus if we add

$$\begin{array}{ll} & 3x^2 + 7x + 8 \\ \text{and} & x^2 + 4x + 5 \\ \hline \text{we get} & 4x^2 + 11x + 13, \end{array}$$

and we are content with the coefficients 11 and 13; but if we add

$$\begin{array}{ll} & 3 \cdot 10^2 + 7 \cdot 10 + 8 \\ \text{and} & 10^2 + 4 \cdot 10 + 5 \\ \hline \text{we get} & 4 \cdot 10^2 + 11 \cdot 10 + 13 \end{array}$$

and the coefficients 11 and 13, being larger than 9, have to be altered as above in order to conform with our positional notation.

We may treat decimal fractions in the same way, recalling that 28.345 means

$$2 \cdot 10 + 8 + \frac{3}{10} + \frac{4}{10^2} + \frac{5}{10^3}.$$

Decimal fractions did not come into use until the time of the French Revolution. In America and England systems of measurement are still used which do not accommodate themselves well to decimal fractions.

EXERCISES

Perform the following operations first by the mechanical rule and then by deriving the result from our postulates, justifying each step as above.

1. Add 305 and 208.

2. Add 246 and 354.

3. Add 99 and 133.

4. Multiply 27 by 49.

5. Multiply 53 by 26.

6. Multiply 102 by 34.

7. Subtract 335 from 412. Find the steps in the postulational derivation where "borrowing" takes place.

8. Add 32.545 and 49.78.

9. Add 2.47 and 37.73.

10. Multiply 2.8 by 3.6.

11. Multiply .023 by 3.2.

47. Other systems of notation. We have just seen that the convenience in the manipulation of arabic numerals is due to the positional character of the notation, which makes possible the use of only 10 digits and such devices as carrying. The question arises whether or not one might have such systems of positional

notation which are not *decimal*, that is, not based on powers of 10. It has been suggested that our system is decimal because of the fact that we have 10 fingers on which we first learned to count. Thus we use the ten digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and we write twenty-four as 24, meaning "two tens and four units." Suppose, however, that we were all born with only one hand with five fingers. Then it is likely that we would write our numbers in powers of five, and twenty-four would be written as 44, meaning "four fives and four units" or $4 \cdot 5 + 4$ instead of 24 or $2 \cdot 10 + 4$. In this system we would have only five digits, 0, 1, 2, 3, 4. That is, we would use a units column, a fives column, a five-squared column, and so on, instead of a units column, a tens column, a ten-squared column, and so on. We would then write:

	one	*two*	*three*	*four*	*five*	*six*	*seven*	*eight*	*nine*	*ten*	*eleven*	*twelve*	
in the ten scale	1	2	3	4	5	6	7	8	9	10	11	12	...
in the five scale	1	2	3	4	10	11	12	13	14	20	21	22	...

The symbols 14 and 22, for example, in the last line should be read as one-four and two-two, respectively, and not fourteen and twenty-two, since they mean "one five and four units" or $1 \cdot 5 + 4$, and "two fives and two units" or $2 \cdot 5 + 2$, respectively.

For example, the number written as 23 in the 10 scale means $2 \cdot 10 + 3$ which is equal to $4 \cdot 5 + 3$ and is therefore written as 43 in the 5 scale.

Example. A number is written as 38 in the ten scale; rewrite it in the five scale. Clearly the highest power of five which is contained in 38 is 5^2. There are one 5^2, two fives, and three units in thirty-eight. That is, $38 = 1 \cdot 5^2 + 2 \cdot 5 + 3$. Hence, thirty-eight is written as 123 in the five scale.

Other scales could be used equally well. In writing a number in the 3 scale, we use only the three digits, 0, 1, 2. Hence twenty-four is written in the 3 scale as 220 since $24 = 2 \cdot 3^2 + 2 \cdot 3 + 0$. In the two scale, we use only the two digits 0, 1. Hence twenty-four is written as 11000 in the two scale, since $24 = 1 \cdot 2^4 +$

$1 \cdot 2^3 + 0 \cdot 2^2 + 0 \cdot 2 + 0$. The same number is written in different ways if we use different scales of notation.

Changing from one scale to another is merely a problem for the packing department. When thirty-eight is written in the 10-scale, it is as though thirty-eight objects were packed into 3 packages of 10 each and 8 packages of one each ($38 = 3 \cdot 10 + 8$). Changing to the 5-scale amounts to repacking the thirty-eight objects into one package of $25 (= 5^2)$, 2 packages of 5 each and 3 packages of 1 each ($123 = 1 \cdot 5^2 + 2 \cdot 5 + 3$). In the 10-scale we pack by ones, 10's, 10^2's, 10^3's and so on. In the 5-scale, we pack by ones, 5's, 5^2's, 5^3's and so on.

It must be emphasized that we are not discussing a different algebra or a different number system here as we were in section 44; we are merely discussing different ways of writing the same old numbers. That is, we are discussing different systems of notation or different (written) languages. When we say that the number which is written as 24 in our usual decimal language would be written as 44 in the "5-scale" language, we are merely translating from one language to another. In the following table, each horizontal line contains just one number rewritten in several different languages.

10 scale	*5 scale*	*3 scale*	*2 scale*	*7 scale*	*9 scale*
24	44	220	11000	33	26
31	111	1011	11111	43	34
8	13	22	1000	11	8

Similarly the same symbol, say 231, would mean different things according to the system of notation in which it is intended to be interpreted. In the usual 10 scale, the symbol 231 means $2 \cdot 10^2 + 3 \cdot 10 + 1$, or the value of the polynomial $2x^2 + 3x + 1$ when $x = 10$, or two hundred and thirty-one. In the 5 scale, 231 means $2 \cdot 5^2 + 3 \cdot 5 + 1$, or the value of the polynomial $2x^2 + 3x + 1$ when $x = 5$, or the number sixty-six.

These new notations look strange only because we have been brought up since early childhood to interpret numerals automatically in the 10 scale. The processes of adding and multiplying would be quite as easy in another scale, say the five scale, including the device of "carrying"; only we would find it clumsy

unless we memorized new tables of addition and multiplication for the other scale as thoroughly as we have memorized those of the 10 scale. In the 5 scale, for example, we have the addition table:

$1 + 1 = 2$	$2 + 1 = 3$	$3 + 1 = 4$	$4 + 1 = 10$
$1 + 2 = 3$	$2 + 2 = 4$	$3 + 2 = 10$	$4 + 2 = 11$
$1 + 3 = 4$	$2 + 3 = 10$	$3 + 3 = 11$	$4 + 3 = 12$
$1 + 4 = 10$	$2 + 4 = 11$	$3 + 4 = 12$	$4 + 4 = 13.$

To add numbers like 34 and 42 (both are written in the 5 scale) we proceed as follows

$$\begin{array}{r} 1 \\ 34 \\ 42 \\ \hline 131. \end{array} \tag{1}$$

That is $2 + 4$ from the above table is 11. We write the right-hand 1 and "carry" the other; then $4 + 4 = 13$. This may be justified by our axioms just as for the 10 scale. For, in the 5 scale, we have $34 = 3 \cdot 5 + 4$ and $42 = 4 \cdot 5 + 2$. Thus

$$\begin{aligned} &(3 \cdot 5 + 4) + (4 \cdot 5 + 2) = 3 \cdot 5 + 4 + 4 \cdot 5 + 2 \\ &= 3 \cdot 5 + 4 \cdot 5 + 4 + 2 = 3 \cdot 5 + 4 \cdot 5 + 4 + 1 + 1 \\ &= 3 \cdot 5 + 4 \cdot 5 + 1 \cdot 5 + 1 = (3 + 4 + 1) \cdot 5 + 1 \\ &= 8 \cdot 5 + 1 = (5 + 3) \cdot 5 + 1 = 1 \cdot 5^2 + 3 \cdot 5 + 1 \\ &= 131 \text{ in the 5 scale.} \end{aligned}$$

The reader should justify each step. The example (1) rewritten in the 10 scale would look like this:

$$\begin{array}{r} 1 \\ 19 \\ 22 \\ \hline 41. \end{array} \tag{2}$$

That is, *(1) and (2) represent exactly the same numerical facts; they are merely written in different languages.* Numbers written in the 10, 5, 3, or 2 scale are often said to be written in the **decimal, quinary, ternary,** or **binary scale,** respectively.

We can also use scales larger than 10 if we invent new digits. For example, in the 12 scale we would have twelve digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, t, e where t and e are digits representing ten and eleven respectively. In the 12 scale we would write twelve

as 10 or $1 \cdot 12 + 0$, thirteen as 11 or $1 \cdot 12 + 1$, one hundred and seventy-one as 123 or $1 \cdot 12^2 + 2 \cdot 12 + 3$, and two hundred and seventy-five as $1te$ or $1 \cdot 12^2 + t \cdot 12 + e$ or one "twelve-squared" plus ten "twelves" plus eleven units (one gross, ten dozen, and eleven units). In fact the Babylonians, thousands of years ago, used a positional system of notation based on 60.

There is little reason to prefer one of these scales to another except that we are familiar with the ten scale. In some parts of higher mathematics it is, however, quite convenient to use other scales, notably the ternary a

E:

Write each of the following number

(*a*) *the 10 scale;* (*b*)
(*d*) *the 8 scale;* (*e*)
(*g*) *the 4 scale;* (*h*)

1. twenty-four. **2.**
4. five. **5.**

7. What number would be repres
to be interpreted in (*a*) the 5 scale; (*b*
Write your answers in the usual 10 s

8. If the number 222 is already v
scale; (*b*) the 2 scale.

9. Rewrite the entire statement
this time expressing it in (*a*) the 5 sc
scale; (*e*) the 8 scale.

10. Add 234 and 432 (both alread
table in the text. Translate both nu
check.

11. Make up a table of multiplicat
by 34 (both already written in the 5
product into the 10 scale and check.

12. Make up a table of addition fo
(both already written in the 3 scale).
scale and check.

13. Make up a table of multiplication for the 3 scale and use it to multiply 102 by 22 (both already written in the 3 scale). Translate both numbers and their product into the 10 scale and check.

14. The following numbers are already written in the twelve scale; rewrite them in the 10 scale:

(*a*) 23; (*b*) $t\,1$; (*c*) $t\,0\,e$; (*d*) $2\,et$.

15. The following numbers are already written in the 10 scale; rewrite them in the 12 scale:

(*a*) 64; (*b*) 49; (*c*) 120; (*d*) 131; (*e*) 167.

48. Further progress in arithmetic. We have already seen the great simplifications introduced into arithmetic by the arabic notation and by the use of decimal fractions. However, in the early years of the 17th century, astronomers like Tycho Brahe and Galileo had made extensive observations and Kepler and others were attempting to reason about heavenly bodies on the basis of these observations. Now in astronomical work the distances involved are very large while the angles are often very small and the calculations become unbearably tedious. Astronomers found that they were literally wasting years of their lives on computation. In response to the need for still easier methods of computation, two men, Napier (about 1614), a Scotsman, and Bürgi, a Swiss, independently discovered a further great simplification of arithmetic called logarithms. We shall be unable to discuss here the way in which these men actually were led to the idea of logarithms since their methods were too advanced for this book. However, we shall introduce logarithms by a method developed later, which is much easier to follow; that is, by way of exponents. First, however, we shall have to extend the notion of exponent, much as we extended the concept of number itself in Chapters III and IV, beginning with the positive integers and going on to other kinds of numbers.

49. Positive integral exponents. We already have the following definition of exponent.

DEFINITION 1. *If n is a positive integer then $x^n = xx \cdots x$ where there are n factors on the right.*

The restriction to positive integers is necessary because only positive integers are used for counting and we have defined the exponent here as the number of factors. We can now prove the following theorems, *in which p and q are understood to be positive integers.*

THEOREM I. $x^p x^q = x^{p+q}$.

Proof. See section 30.

THEOREM IIa. *If $p > q$, $x^p/x^q = x^{p-q}$ $(x \neq 0)$.*

Proof. $$x^p = (xx \cdots x)(xx \cdots x)$$
where the first parenthesis has $p - q$ factors and the second has q factors. Hence,

$$\frac{x^p}{x^q} = \frac{\overbrace{(x \cdot x \cdots x)}^{(p - q \text{ factors})}\overbrace{(x \cdot x \cdots x)}^{(q \text{ factors})}}{\underbrace{(x \cdot x \cdots x)}_{(q \text{ factors})}} = \overbrace{(x \cdot x \cdots x)}^{(p - q \text{ factors})}$$
$$= x^{p-q}$$

by definition. The restriction $x \neq 0$ must be made because if $x = 0$ the expression x^p/x^q becomes $0/0$ which has been excluded.

For example, $\dfrac{x^5}{x^2} = \dfrac{(xxx)(xx)}{(xx)} = xxx = x^3 = x^{5-2}$.

THEOREM IIb. *If $p < q$, $x^p/x^q = 1/x^{q-p}$ $(x \neq 0)$.*

Proof. $x^q = (x \cdot x \cdots x)(x \cdot x \cdots x)$ where the first parenthesis has $q - p$ factors and the second has p factors. Hence,

$$\frac{x^p}{x^q} = \frac{\overbrace{(x \cdot x \cdots x)}^{(p \text{ factors})}}{\underbrace{(x \cdot x \cdots x)}_{(q - p \text{ factors})}\underbrace{(x \cdot x \cdots x)}_{(p \text{ factors})}} = \frac{1}{\underbrace{(x \cdot x \cdots x)}_{(q - p \text{ factors})}} = \frac{1}{x^{q-p}}.$$

For example, $\dfrac{x^2}{x^5} = \dfrac{(xx)}{(xxx)(xx)} = \dfrac{1}{(xxx)} = \dfrac{1}{x^3} = \dfrac{1}{x^{5-2}}$.

THEOREM IIc. *If $p = q$, then $x^p/x^q = 1$ $(x \neq 0)$.*

Proof. $x^p/x^q = x^p/x^p = 1.$

For example, $x^3/x^3 = 1.$

THEOREM III. $(x^p)^q = x^{pq}.$

Proof. We have $(x^p)^q = (x^p)(x^p) \cdots (x^p)$, where there are q such parentheses. Thus,

$$(x^p)^q = (x \cdot x \cdots x)(x \cdot x \cdots x) \cdots (x \cdot x \cdots x) = x^{pq},$$

since there are p x's in each of the q parentheses.

For example, $(x^2)^3 = (x^2)(x^2)(x^2) = (xx)(xx)(xx) = x^6 = x^{2 \cdot 3}.$

THEOREM IV. $(xy)^p = x^p y^p.$

Proof. See section 23.

EXERCISES

Simplify:

1. $\dfrac{a^{10}}{a^2}.$ 2. $\dfrac{a^2b^3c^2}{a^5bc^2}.$ 3. $\dfrac{(a^3b)^2}{(ab^2)^3}.$ 4. $(a^2b)(ab^3).$

Evaluate:

5. $2^3 \cdot 3^2.$ 6. $8^9/8^7.$ 7. $(-3)^2 \cdot (-2)^3.$
8. $8^4/2^{11}.$ 9. $2 \cdot 3^2.$ 10. $5 + 2 \cdot 3^2.$

Correct the right member of each of the following, if necessary:

11. $3^2 + 3^3 = 3^5.$ 12. $3^2 + 3^3 = 6^5.$ 13. $3^2 \cdot 3^3 = 3^6.$
14. $3^2 \cdot 3^3 = 9^5.$ 15. $3^8/3^2 = 3^4.$ 16. $(3^4)^2 = 3^6.$
17. $6^3/2^3 = 3^3.$

50. Negative integers and zero as exponents. The occurrence of three separate cases in Theorem II, section 49, is a source of inconvenience and it is natural to desire to unify them.

Suppose we attempted to apply the formula of Theorem IIa,

$$\frac{x^p}{x^q} = x^{p-q},$$

mechanically even if $p < q$ in spite of the fact that it does not apply to this case. This would make the exponent $p - q$ negative. For example,

$$\frac{x^2}{x^5} = x^{2-5} = x^{-3}.$$

But an exponent -3 has no sense, since our definition of exponent is necessarily restricted to positive integers as we have seen. The symbol x^{-3} cannot possibly be interpreted according

to Definition 1. No one can count out -3 factors, except possibly at a séance; but x^{-3} is not intended to mean the product of the ghosts of three departed x's, or something equally mysterious. The symbol x^{-3} does not come under the jurisdiction of our definition at all and is at present a completely undefined symbol. This is fortunate for since x^{-3} has no preassigned meaning we are free to choose a meaning for it to suit ourselves without fear of contradicting previous results. We see at once that we will be able to dispense with Theorem IIb if we adopt the following definition.

DEFINITION 2. *If* $-n$ *is a negative integer* $x^{-n} = \dfrac{1}{x^n} (x \neq 0)$.

Writing $x^2/x^5 = x^{2-5} = x^{-3}$ would now agree with the facts since x^{-3} means $1/x^3$. For example, $2^{-3} = 1/2^3 = 1/8$.

Similarly, if we try to apply the formula of Theorem IIa mechanically even if $p = q$, although it does not apply to this case either, we would get

$$\frac{x^p}{x^q} = x^{p-q} = x^0$$

since $p = q$. But the symbol x^0 has no sense, under definitions 1 or 2. Definition 1 defines positive integral exponents and definition 2 defines negative integral exponents; but 0 is neither positive nor negative. Certainly x^0 cannot mean the result obtained by taking no x's, in some occult way, and multiplying them together. Since the symbol x^0 does not come under the jurisdiction of definitions 1 and 2, we are free to choose any meaning we please for it. We see that we can dispense with Theorem IIc if we adopt the following definition.

DEFINITION 3. $x^0 = 1,\ (x \neq 0)$.

For example, writing $x^3/x^3 = x^{3-3} = x^0$, according to the rule of Theorem IIa, is in agreement with the fact that $x^3/x^3 = 1$, for x^0 now means 1. For example, $7^0 = 1$.

By adopting these supplementary definitions 2 and 3, we extend the concept of exponent to allow all integers, positive, negative, or zero, to occur as exponents. At the same time the three parts of Theorem II reduce to one formula:

$$\frac{x^p}{x^q} = x^{p-q}.$$

Furthermore, this formula in turn reduces to that of Theorem I since if we apply the rule of Theorem I mechanically to x^p/x^q we get

$$\frac{x^p}{x^q} = x^p \cdot \frac{1}{x^q} = x^p x^{-q} = x^{p+(-q)} = x^{p-q},$$

which is in agreement with the facts. Thus, *by adopting our supplementary definitions of negative integral and zero exponents we dispense with Theorem II. entirely.*

Remark. Note that $2^2 = 4$ is half of $2^3 = 8$, and $2^1 = 2$ is half of $2^2 = 4$. By our definitions, $2^0 = 1$ is half of $2^1 = 2$, $2^{-1} = \dfrac{1}{2}$ is half of $2^0 = 1$, and $2^{-2} = \dfrac{1}{4}$ is half of $2^{-1} = \dfrac{1}{2}$. Hence our definitions fit well with our desire for uniformity in the sequence 2^3, 2^2, 2^1, 2^0, 2^{-1}, 2^{-2}, and so on.

It remains to prove that, for any integral exponents p and q, positive, negative, or zero, the three rules

$$x^p x^q = x^{p+q} \tag{1}$$

$$(x^p)^q = x^{pq} \tag{2}$$

$$(xy)^p = x^p y^p \tag{3}$$

still hold. For example, let us prove (3).

Proof. Case 1. Suppose p is positive. Then it has already been proved in section 49.

Case 2. Suppose p is negative. Then $p = -a$ where a is positive. And

$$(xy)^p = (xy)^{-a} = \frac{1}{(xy)^a} \qquad \text{by Definition 2.}$$

But $(xy)^a = x^a y^a$ by case 1, since a is positive. Therefore

$$(xy)^p = \frac{1}{x^a y^a} = \frac{1}{x^a} \cdot \frac{1}{y^a} = x^{-a} y^{-a} = x^p y^p$$

which was to be proved.

Case 3. Suppose $p = 0$. Then $(xy)^p = 1$, $x^p = 1$, and $y^p = 1$, by definition 3. Thus formula (3) is verified immediately for this case since $1 = 1 \cdot 1$.

The proofs of (1) and (2) are included in the following list of exercises.

EXERCISES

Evaluate:

1. 3^{-2}.
2. $\left(\frac{1}{2}\right)^0$.
3. $\left(\frac{1}{2}\right)^{-3}$.
4. $(-2)^{-3}$.
5. $10^0 \cdot 10^3 \cdot 10^{-2}$.
6. $\dfrac{3^2 \cdot 2^3}{3^{-1} \cdot 2^{-1}}$.
7. $\dfrac{10^5 \cdot 10^{-3}}{10^2}$.
8. $\dfrac{10^4 \cdot 10^{-5}}{10^{-3}}$.
9. $(3 \cdot 10^{-1})^2$.
10. $\dfrac{(3 \cdot 10^5)(4 \cdot 10^{-3})}{2 \cdot 10^3}$.
11. $\dfrac{(2 \cdot 10^2)^3 \cdot (4 \cdot 10^{-2})^2}{(2 \cdot 10^3)^{-1}}$.
12. $(10^{-2})^{-3}$.

Simplify the following by removing zero and negative exponents:

13. $3\,x^{-3}$.
14. $(3\,x)^{-3}$.
15. $3\,x^0$.
16. $(3\,x)^0$.
17. $(2\,a^0b^2)^3 \cdot (b^{-2})^2 \cdot (b^3)^0$.
18. $(3\,x^2y^3) \cdot (2\,x^{-3}y^{-2})$.
19. $\dfrac{2\,x^3y^2}{x^{-3}y^{-2}}$.
20. $\dfrac{a^{-1} + b^{-1}}{a^{-1} - b^{-1}}$.
21. $x^{-2}(x^5 + 3\,x^3 + 5\,x^2)$.
22. $\dfrac{(ab^2)^{-3}}{(a^2b)^{-2}}$.
23. $\dfrac{(xy^2)^{-3}(x^{-2}y^{-1})^2}{(x^4y^3)^{-1}}$.
24. $(x^{-2})^{-5}$.

25. Prove that formula (1) holds for *all* integral exponents p and q. (Hint: divide the proof into nine separate cases as follows. Case 1a: p positive, q positive. Case 1b: p positive, q negative. Case 1c: p positive, $q = 0$. Case 2a: p negative, q positive. Case 2b: p negative, q negative. Case 2c: p negative, $q = 0$. Case 3a: $p = 0$, q positive. Case 3b: $p = 0$, q negative. Case 3c: $p = 0$, $q = 0$. Imitate the proof of formula (3) in the text as far as possible).

26. Prove that formula (2) holds for *all* integral exponents p and q. (See hint for exercise 25.)

51. Computation with powers of ten. When very large or very small numbers are used in scientific writing it is customary to express them in terms of powers of ten, as follows. Instead of 1,000,000 one would write 10^6. Similarly 378,000,000 might be written as 378×10^6 or 37.8×10^7 or 3.78×10^8; we use the $\times$ sign for multiplication here to avoid confusion with the decimal point. Similarly, .0001 would be written as 10^{-4} and .000378 would be written as 378×10^{-6} or 37.8×10^{-5} or 3.78×10^{-4}. Note that multiplying a number by 10^4 moves the decimal point to the right four places while multiplying by 10^{-4} moves the decimal point to the left four places.

Every positive real number can be expressed as a number between one and ten multiplied by a suitable integral power of ten; when this is done the number is said to be expressed in ***standard form.*** For

example, $378{,}000{,}000 = 3.78 \times 10^8$, $.000378 = 3.78 \times 10^{-4}$ and $3.78 = 3.78 \times 10^0$. This so-called "standard form" is more compact and more comprehensible than the everyday way of writing such numbers with large numbers of zeros before or after the decimal point. It also lends itself to calculations with cumbersome numbers, large or small, which are made easy by the use of the laws of exponents.

Example. Calculate the value of $\dfrac{378{,}000{,}000{,}000 \times .000004}{2000}$.

We write $\dfrac{3.78 \times 10^{11} \times 4 \times 10^{-6}}{2 \times 10^3} = 3.78 \times 2 \times 10^{11-6-3} =$ $7.56 \times 10^2 = 756$. The coefficients of the powers of ten have to be multiplied and divided in the ordinary way. How this can be avoided by expressing them too as (fractional) powers of ten, will be discussed in sections 54 and 55.

EXERCISES

Express each of the following in standard form:

1. 200,000. **2.** 3,400,000. **3.** 567,000,000,000. **4.** .3.
5. .056. **6.** .000782. **7.** .0000000893. **8.** 3.
9. 46.78.

Express each of the following in ordinary positional (decimal) notation:

10. 10^5. **11.** 10^{-1}. **12.** 10^{-3}. **13.** 6.78×10^7.
14. 6.78×10^{-5}. **15.** 6.78×10^0. **16.** 3.76×10^{-4}. **17.** 4.68×10^5.
18. 46.7×10^{-4}.

Calculate, using the laws for exponents and express the result in ordinary positional (decimal) notation:

19. $\dfrac{(6 \times 10^7)\times(3 \times 10^{-3})}{2 \times 10^3}$. **20.** $\dfrac{(5.2 \times 10^6)\times(4 \times 10^{-8})}{13 \times 10^{-4}}$

21. If c represents the velocity of radiant energy in a vacuum, L its wave length, and f its frequency, then $f = c/L$. Suppose c is 3×10^{10} centimeters per second and $L = 6 \times 10^{-5}$ cm. Find f. (Hint: write $3 \times 10^{10} = 30 \times 10^9$.)

22. The angstrom is a unit of length equal to 10^{-8} centimeters. If one centimeter is .3937 inches, express in inches the wave length of red light with a wave length of 8000 angstroms.

23. A light year is a unit of length equal to 5.88×10^{12} miles. The distance of the cluster of stars called the Pleiades is 1.2936×10^{15} miles. How many light years is this?

24. The Great Nebula in Andromeda is approximately 5.292×10^{18} miles away. How many light years is this? Use the data of exercise 23.

25. If a gram is .002205 pounds and the mass of the earth is 5.97×10^{27} grams. Find the mass of the earth in pounds.

26. The mass of the sun is 1.98×10^{33} grams. Find the mass of the sun in pounds. Use the data of exercise 25.

52. Fractional exponents. Since we have successfully extended the concept of exponent to include all integers as exponents, it is natural to ask whether we cannot extend the notion still further, say to fractions. In making such an extension, since if we do it at all we do it for our own convenience, we shall want to preserve the validity of formulas (1), (2), (3) of section 50; just as, when extending the notion of number itself in Chapters III and IV we desired to preserve the validity of such formulas as the associative, commutative and distributive laws.

Let us experiment with a symbol like $x^{\frac{1}{2}}$. This has no meaning under any of the previous definitions. It certainly cannot mean that we take a product of x's, the number of factors being 1/2. In particular, $X^{\frac{1}{2}}$ does not mean λ. Since $x^{\frac{1}{2}}$ has no previous meaning, we are free to choose any definition that suits us, and it will suit us to give such a definition, if possible, as will preserve the validity of our 3 formulas. Consider formula (2). If this is to operate, then $x^{\frac{1}{2}}$, whatever it shall mean, must satisfy the relation $(x^{\frac{1}{2}})^2 = x^{\frac{1}{2} \cdot 2} = x^1 = x$. Thus $x^{\frac{1}{2}}$ will have to mean something which when squared yields x. But this can only be $\sqrt{x}$ or $-\sqrt{x}$. Therefore we choose the *definition:* $x^{\frac{1}{2}} = \sqrt{x}$. For example, $9^{\frac{1}{2}} = 3$. This obviously fits in with formula (1) as well, since $x^{\frac{1}{2}} \cdot x^{\frac{1}{2}} = x^{\frac{1}{2}+\frac{1}{2}} = x^1 = x$. More generally we choose the definition $x^{\frac{1}{p}} = \sqrt[p]{x}$, where p is a positive integer, so that rule (2) continues to operate as follows:

$$(\sqrt[p]{x})^p = (x^{\frac{1}{p}})^p = x^{\frac{1}{p}\cdot p} = x^1 = x.$$

Furthermore if rule (2) is to remain valid we would have

$$(\sqrt[q]{x})^p = (x^{\frac{1}{q}})^p = x^{\frac{1}{q}\cdot p} = x^{\frac{p}{q}}, \qquad \text{and} \qquad \sqrt[q]{x^p} = (x^p)^{\frac{1}{q}} = x^{p\cdot\frac{1}{q}} = x^{\frac{p}{q}}$$

Hence we are led to make the following general definition.

DEFINITION 4. $x^{\frac{p}{q}} = (\sqrt[q]{x})^p = \sqrt[q]{x^p}$, *where* p *and* q *are any integers* $(q > 0)$.

Any rational number can be written in the form p/q where p and q are integers and $q > 0$, that is, with a positive denomi-

nator. For example, $3/-4$ can be written as $-3/4$ and $-5/-6$ can be written as $5/6$. Definition 4 therefore extends the notion of exponent to *all* rational numbers. As in section 23, we confine ourselves to roots of positive numbers. Thus, in the present section, the symbols x, a, b represent positive numbers.

For example, $x^{\frac{2}{3}} = \sqrt[3]{x^2} = (\sqrt[3]{x})^2$. Thus $8^{\frac{2}{3}} = (\sqrt[3]{8})^2 = 2^2 = 4$. Similarly, $8^{-\frac{2}{3}} = \sqrt[3]{8^{-2}} = \sqrt[3]{1/8^2} = \sqrt[3]{1/64} = 1/4$.

We must not take it for granted that $x^{\frac{2}{3}}$ means the same thing as $x^{\frac{4}{6}}$, say. This has to be proved. For $x^{\frac{2}{3}}$ means a cube root of x^2, that is, a number a such that $a^3 = x^2$, by definition; while $x^{\frac{4}{6}}$ means a sixth root of x^4. Now, squaring a^3 we get $(a^3)^2 = (x^2)^2$ or $a^6 = x^4$. That is, a *is* a sixth root of x. Hence $x^{\frac{2}{3}} = x^{\frac{4}{6}}$. In the same way we can prove the following general theorem.

THEOREM A. $x^{\frac{p}{q}} = x^{\frac{pr}{qr}}$ *where* p, q, r *are integers and* q *and* qr *are understood to be positive.*

Proof. By definition $x^{\frac{p}{q}}$ means a number a such that $a^q = x^p$ (that is, a qth root of x^p), while $x^{\frac{pr}{qr}}$ means a qr th root of x^{pr}. Raising a^q to the r th power we obtain $(a^q)^r = (x^p)^r$, or $a^{qr} = x^{pr}$. That is, a *is* a qr th root of x^{pr}. This proves the theorem.

We shall need the following theorem.

THEOREM B. *If* n *is any positive integer,* $\sqrt[n]{a} \cdot \sqrt[n]{b} = \sqrt[n]{ab}$.

Proof. See section 23.

We must not assume naively that the formulas (1), (2), (3) of section 50 remain valid for fractional exponents since they have been proved only for integral exponents. However, we are now in a position to prove them for fractional exponents. For example, let us prove that $x^{\frac{1}{2}} \cdot x^{\frac{1}{3}} = x^{\frac{1}{2}+\frac{1}{3}}$ or $x^{\frac{5}{6}}$. Now $x^{\frac{1}{2}} = x^{\frac{3}{6}}$ and $x^{\frac{1}{3}} = x^{\frac{2}{6}}$ by Theorem A. Hence

$$\begin{aligned} x^{\frac{1}{2}} \cdot x^{\frac{1}{3}} &= x^{\frac{3}{6}} \cdot x^{\frac{2}{6}} = \sqrt[6]{x^3} \cdot \sqrt[6]{x^2} && \text{(by definition)} \\ &= \sqrt[6]{x^3 \cdot x^2} && \text{(by Theorem B)} \\ &= \sqrt[6]{x^5} && \text{(by (1) section 50, since 3 and 2 are integers)} \\ &= x^{\frac{5}{6}} && \text{(by definition).} \end{aligned}$$

Similarly we can prove that formula (1) of section 50 holds for all fractional exponents, as follows.

THEOREM C. *$x^{\frac{p}{q}} \cdot x^{\frac{u}{v}} = x^{\frac{p}{q} + \frac{u}{v}}$ or $x^{\frac{pv+qu}{qv}}$ where p, q, u, v are integers and q and v are understood to be positive.*

Proof. By Theorem A, $x^{\frac{p}{q}} = x^{\frac{pv}{qv}}$ and $x^{\frac{u}{v}} = x^{\frac{qu}{qv}}$.

Hence, $x^{\frac{p}{q}} \cdot x^{\frac{u}{v}} = x^{\frac{pv}{qv}} \cdot x^{\frac{qu}{qv}} = \sqrt[qv]{x^{pv}} \cdot \sqrt[qv]{x^{qu}}$ (by definition)

$= \sqrt[qv]{x^{pv} \cdot x^{qu}}$ (by Theorem B)

$= \sqrt[qv]{x^{pv+qu}}$ (by (1), section 50, since pv and qu are integers)

$= x^{\frac{pv+qu}{qv}}$ (by definition).

We have proved that (1) of section 50 holds for *all* rational exponents. By similar methods we could prove that (2) and (3) of section 50 hold for all rational exponents. We shall not carry out the detailed proofs here; they are included in exercise 20 below.

The validity of these formulas for fractional exponents provides many algebraic simplifications. For example, it is not easy at first glance to simplify such an expression as $(\sqrt[3]{x^2})(\sqrt[5]{x})^3$. But, using fractional exponents it becomes

$$x^{2/3} \cdot x^{3/5} = x^{\frac{2}{3}+\frac{3}{5}} = x^{19/15} = \sqrt[15]{x^{19}} \text{ or } (\sqrt[15]{x})^{19}.$$

We have now extended the concept of exponent to allow all rational numbers as exponents. It is possible to extend it further to allow irrational and even imaginary exponents but this will not be done here.

EXERCISES

Evaluate:

1. $16^{1/2}$. **2.** $16^{1/4}$. **3.** $8^{1/3}$. **4.** $8^{-1/3}$. **5.** $16^{-1/2}$.
6. $8^{2/3}$. **7.** $8^{-2/3}$. **8.** $16^{3/4}$. **9.** $16^{-3/2}$. **10.** $8^{-4/3}$.
11. $27^{-5/3}$. **12.** $32^{-3/5}$. **13.** $9^{-1/2}$. **14.** $16^{5/4}$.

Simplify:

15. $\sqrt[4]{x^3} \cdot \sqrt[7]{x^5}$. **16.** $\sqrt[5]{x^2} \cdot (\sqrt[5]{x})^3$. **17.** $\sqrt[5]{x^7}/\sqrt[3]{x^4}$.

18. Evaluate $(27^{2/3} + 8^{-1/3}) \div (8^{3/5} \cdot 8^{2/5})$.

19. Using the definitions, prove, without assuming the validity of formulas (1), (2), and (3) of section 50 for fractional exponents, that:

(*a*) $(x^{1/3})^2 = x^{2/3}$. (*b*) $x^{1/5} \cdot x^{2/5} = x^{3/5}$.
(*c*) $(x^{1/3})^{1/2} = x^{1/6}$. (*d*) $x^{1/4} \cdot x^{2/3} = x^{11/12}$.

20. Prove the following using the definitions, but not assuming the validity of formulas (1), (2), (3) of section 50 for fractional exponents, p, q, u, v being integers, and q and v being positive: (*a*) $(x^{\frac{p}{q}})^{\frac{u}{v}} = x^{\frac{pu}{qv}}$; (*b*) $(xy)^{\frac{p}{q}} = x^{\frac{p}{q}} \cdot y^{\frac{p}{q}}$.

53. Logarithms. While the algebraic simplifications introduced by the use of fractional and negative exponents are important, even more important is the fact that these exponents provide a foundation for the theory of logarithms which, in turn, must be considered one of the major achievements of the 17th century. Logarithms simplified computation to such an extent that it is difficult to estimate the time saved thereby for astronomers and other people who are faced with tedious calculations. It is important to memorize the following fundamental definition.

DEFINITION 5. *If b is any positive number, different from 1, and $b^y = x$, then the exponent y is called the* **logarithm of *x* to the base *b*.** *In symbols, $y = \log_b x$.*

For example, $3 = \log_2 8$ because $2^3 = 8$. Similarly $2 = \log_3 9$ because $3^2 = 9$.

It is easy to confuse the roles played by the three numbers b, y, and x in this definition unless it is learned carefully. Notice that a logarithm *is* an exponent; that is, $\log_b x$ means the exponent y to which we must raise the base b in order to get the number x. It can be proved that *every positive real number x can be expressed as b^y, where y is a real number, rational or irrational.* This will not be done here.

Exercise. Why do we make the restriction, in definition 5, that $b \neq 1$?

Remark. There are reasons for restricting b to be positive. For example, if b were negative, b^2 would be positive, b^3 negative, b^4 positive, and so on; and $b^{\frac{1}{2}}$ would be imaginary. This would introduce technical difficulties which are best avoided here. If b is positive, however, and y is real (positive, negative or zero), x is still positive. We shall confine ourselves to positive x's; that is, we shall speak of $\log_b x$ only when b and x are positive numbers.

How this definition enables us to simplify arithmetical calculations will be made clear in the next section. The simplifications are essentially due to the following theorems.

THEOREM I. $\log_b (xw) = \log_b x + \log_b w$.

THEOREM II. $\log_b (x/w) = \log_b x - \log_b w$.

THEOREM III. $\log_b (x^r) = r \log_b x$.

Since logarithms are nothing but exponents it is natural to expect to prove these theorems by using the laws of operation with exponents. In fact, these three theorems will be seen to be nothing more than the translations (from the language of exponents into the language of logarithms) of the following known laws for exponents:

$$(a)\quad b^y b^u = b^{y+u}$$
$$(b)\quad b^y/b^u = b^{y-u}$$
$$(c)\quad (b^y)^r = b^{yr}.$$

Let us prove our three theorems. Let $y = \log_b x$, and $u = \log_b w$; hence we have $b^y = x$ and $b^u = w$.

Proof of Theorem I. By (a), $xw = b^y b^u = b^{y+u}$. But $xw = b^{y+u}$ says that $\log_b (xw) = y + u$ by definition 5. Substituting for y and u, we get $\log_b (xw) = \log_b x + \log_b w$.

Proof of Theorem II. By (b), $x/w = b^y/b^u = b^{y-u}$. By definition 5 this says that $\log_b (x/w) = y - u = \log_b x - \log_b w$.

Proof of Theorem III. By (c), $x^r = (b^y)^r = b^{yr}$. By definition 5, this says that $\log_b (x^r) = yr = r \log_b x$. (Note: we write $r \log_b x$ to avoid the confusion that would arise from writing $\log_b x \cdot r$ as to whether we meant $\log_b (xr)$ or $(\log_b x)r$.)

EXERCISES

Express in logarithmic notation:

1. $5^2 = 25$. **2.** $3^4 = 81$. **3.** $10^2 = 100$. **4.** $25^{1/2} = 5$.
5. $8^{1/3} = 2$. **6.** $8^{-1/3} = 1/2$. **7.** $5^{-2} = 1/25$. **8.** $2^{-3} = 1/8$.
9. $7^0 = 1$. **10.** $10^{-1} = 1/10$.

Express in exponential notation:

11. $\log_2 16 = 4$. **12.** $\log_{10} 1000 = 3$. **13.** $\log_{10} .01 = -2$.
14. $\log_{10} 10 = 1$. **15.** $\log_{10} 1 = 0$. **16.** $\log_2 8 = 3$.
17. $\log_5 125 = 3$. **18.** $\log_{10} 100 = 2$. **19.** $\log_{10} .1 = -1$.
20. $\log_4 8 = 3/2$.

Find:

21. $\log_2 32$. **22.** $\log_{10} .001$. **23.** $\log_8 4$. **24.** $\log_5 1$.
25. $\log_5 5$. **26.** $\log_b b$. **27.** $\log_b 1$. **28.** $\log_b (b^3)$.
29. $\log_{25} 125$. **30.** $\log_9 27$. **31.** $\log_{10} 10{,}000$.
32. $\log_{10} (10^2)$. **33.** $\log_{10} (10^5)$. **34.** $\log_b (b^7)$.

If $\log_{10} 2 = .3010$ and $\log_{10} 3 = .4771$, use Theorems I, II, III, to find:

35. $\log_{10} 6$. **36.** $\log_{10} 4$. **37.** $\log_{10} (3/2)$.
38. $\log_{10} 9$. **39.** $\log_{10} 12$. **40.** $\log_{10} 20$.
41. $\log_{10} .2$. **42.** $\log_{10} 18$.

54. Common logarithms. Any positive number $b \neq 1$ can be used as a base for logarithms. An irrational number called e whose decimal expression is 2.71828 is actually used in advanced mathematics for theoretical reasons which cannot be explained here. Logarithms with the base e are called **natural logarithms,** a name that must seem very inappropriate to you, and are closely related to the original systems of Napier and Bürgi. An improvement of Napier's system, which is said to have occurred to both Napier and H. Briggs, an English mathematician, led to the system of logarithms to the base 10, known as **common logarithms.** Logarithms to the base 10 are the most convenient for computation simply because we write our numbers in decimal notation.* It is with this system that we work hereafter. We shall write $\log x$, without any base indicated, to mean $\log_{10} x$. Thus we write $\log 100 = 2$ because $10^2 = 100$; $\log 1000 = 3$ because $10^3 = 1000$, $\log .1 = -1$, because $10^{-1} = 1/10 = .1$, and $\log 1 = 0$ because $10^0 = 1$. A rudimentary table of common logarithms is easy to construct:

Table A

x	$\cdots$	.0001	.001	.01	.1	1	10	100	1000	10000	$\cdots$
$\log x$	$\cdots$	-4	-3	-2	-1	0	1	2	3	4	$\cdots$

Since 34.2 is between 10 and 100 it is natural to expect $\log 34.2$ to be between $\log 10 = 1$ and $\log 100 = 2$. That is, in exponential language, since $10^1 = 10$ and $10^2 = 100$ it is natural to expect 10 raised to some exponent between 1 and 2 to be exactly 34.2. This is actually so, but the proof is too difficult to be given here. It can be shown that

$$\log 34.2 = 1.5340$$

approximately; that is, in exponential notation,

* If we were in the habit of writing numbers in the 5 scale, then logarithms to the base 5 would be most convenient for computation.

$$10^{1.5340} = 34.2 \text{ or } 10^{1534/1000} = 34.2$$

approximately; in radical notation

$$\sqrt[1000]{10^{1534}} = 34.2$$

approximately.

We have remarked, in section 51, that every positive number can be expressed as the product of a number between 1 and 10 and an integral power of 10. For example, $34.2 = 3.42(10)$. Similarly,

$$\begin{aligned} 534 &= 5.34(10^2), \\ .00534 &= 5.34(10^{-3}), \end{aligned}$$

and

$$5.34 = 5.34(10^0).$$

By Theorem I, section 53 we have

$$\begin{aligned} \log 534 &= \log 5.34 + \log (10^2), \\ \log .00534 &= \log 5.34 + \log (10^{-3}), \\ \log 5.34 &= \log 5.34 + \log (10^0). \end{aligned}$$

By the definition of logarithm, or by Table A, above, we obtain

$$\begin{aligned} \log 534 &= \log 5.34 + 2 \\ \log .00534 &= \log 5.34 + (-3) \\ \log 5.34 &= \log 5.34 + 0. \end{aligned}$$

Therefore we see that we could write down the logarithm of any number if we knew only *the logarithms of numbers from 1 to 10.* The latter are called **mantissas** and a table of mantissas approximated to 4 decimal places (Table I) can be found at the end of this book. They are decimals since the logarithm of a number between 1 and 10 must be between 0 and 1.

From the tables, we obtain $\log 5.34 = .7275$. Hence,

$$\log 534 = 2.7275$$

and

$$\log .00534 = -3 + .7275 = -2.2725.$$

Notice that since the digits of the numbers 5.34, 534, and .00534 are the same their logarithms have the same decimal part (.7275) or mantissa, because any of these numbers can be obtained from any other of them by multiplying by a suitable integral power of 10. Therefore, their logarithms differ only by a whole number. *When a logarithm is expressed so that its decimal part is written positively, the integral part of the logarithm is called its* ***charac-***

teristic. For example, the characteristic of log .00534 is -3, not -2 (see above). The above discussion shows us that *to find the logarithm of a number we have only to estimate its characteristic from Table A * and look up its mantissa in Table I*. For example, log 534 is between 2 and 3 since 534 is between $100 = 10^2$ and $1000 = 10^3$ (see Table A). Hence, the characteristic is 2. The mantissa, as we know from Table I, is .7275. Hence log 534 = 2.7275. Similarly, log .00534 is between -3 and -2 because .00534 is between $.001 = 10^{-3}$ and $.01 = 10^{-2}$ (Table A). Hence, $\log .00534 = -3 + .7275$. Note that log .00534 is *not* -3.7275; this would be between -3 and -4 instead of between -3 and -2. We usually find it more convenient to write logarithms with negative characteristics, like the latter, in a different way. Instead of writing $\log .00534 = -3 + .7275$ we borrow a trick from the politicians and give 10 with our left hand while taking 10 away with our right, as follows:

$$\log .00534 = 10 - 3 + .7275 - 10 = 7.7275 - 10.$$

The latter form is less clumsy in practice.

The tables may also be used to find a number whose logarithm is known. If we know that $\log x = 3.7275$ we look up .7275 in the table of mantissas and find that x has the digits 534. The characteristic 3 tells us that x is between $10^3 = 1000$ and $10^4 = 10000$ (Table A) since $\log x$ is between 3 and 4. Hence $x = 5340$, approximately. Similarly, if $\log x = 8.7275 - 10$ the digits are still 534 but the characteristic $8 - 10 = -2$ tells us that x is between $10^{-2} = .01$ and $10^{-1} = .1$ since $\log x$ is between -2 and -1 (Table A). Hence $x = .0534$, approximately.

If we do not find a mantissa in the table exactly we will take the nearest one to it, since we are interested only in approximate results. If more accuracy is desired one can use tables computed

* Various rules are often stated for determining the characteristic. For example: if $N \geqq 1$ then the characteristic of $\log N$ is one less than the number of digits in N to the left of the decimal point; but if $N < 1$, and if the first non-zero digit of N is in the k th decimal place, then $-k$ is the characteristic of $\log N$. One may also state the rule as follows: express the number N in "standard form," that is, as the product of a number between 1 and 10 and an integral power of ten; then the exponent of ten is the characteristic of $\log N$. Or one may state it as follows: move the decimal point to where it would be if we were to write N in "standard form"; if we moved it k places to the left, the characteristic of $\log N$ is k, and if we moved it k places to the right, the characteristic is $-k$. None of these rules need be memorized if one thinks of Table A.

for more than 3 digits and more than 4 decimal places. Another device for obtaining further accuracy from a given table, known as *interpolation,* will be discussed in Chapter X, section 94.

Notice that while Table I is usually called a table of common logarithms, it is really only a table of mantissas. How these tables are computed will be discussed briefly in Chapter XI, section 115.

EXERCISES

Find:

1. log 124. **2.** log 12.4. **3.** log .0124. **4.** log 387.
5. log 39.6. **6.** log 5.34. **7.** log .0435. **8.** log .467.
9. log .000346. **10.** log 306. **11.** log 360. **12.** log 500.
13. log .005.

Find the number whose logarithm is:

14. 2.5211. **15.** 1.5211. **16.** 8.5211 − 10. **17.** 9.5211 − 10.
18. 0.5211. **19.** 1.6972. **20.** 0.7316. **21.** 9.7364 − 10.
22. 8.9400 − 10. **23.** 3.9717. **24.** 2.5888.

25. Express the statement log 862 = 2.9355 in exponential notation; in radical notation.

26. Express the statement log 631 = 2.8000 in exponential notation; in radical notation.

55. Computation with logarithms. Having learned to use the tables, the simplification in computation is obtained by the use of Theorems I, II, III of section 53. Two illustrations will suffice to make the method clear.

Example 1. Calculate $\dfrac{(8.34)(65.2)}{376}$.

Call the result x. Then by Theorems I and II,

$$\log x = \log \frac{(8.34)(65.2)}{376} = \log 8.34 + \log 65.2 - \log 376.$$

Now

$$\begin{aligned} \log 8.34 &= 0.9212 \\ \log 65.2 &= \underline{1.8142} \\ \log 8.34 + \log 65.2 &= 2.7354 \\ \log 376 &= \underline{2.5752} \\ \log x &= 0.1602. \end{aligned}$$

Hence $x = 1.45$, approximately.

Example 2. Find $\sqrt[3]{473}$. Let x be the result. Then

$$x = \sqrt[3]{473} = 473^{1/3}.$$

By Theorem III, $\log x = \frac{1}{3}\log 473 = \frac{1}{3}(2.6749) = 0.8916.$

Hence $x = 7.79$, approximately.

Example 2 has been done far more quickly than would be possible by the method of section 25. Even the relatively simple example 1 has been done more quickly than by straightforward arithmetic. This is so essentially because multiplication is replaced by addition of logarithms, division by subtraction of logarithms and the extraction of roots by simple division of a logarithm by the index of the root. What we have done with our table is to represent every positive number approximately as a power of 10. In example 1, we have

$$8.34 = 10^{0.9212},\ 65.2 = 10^{1.8142},\ 376 = 10^{2.5752}$$

and hence
$$\frac{(8.34)(65.2)}{376} = \frac{(10^{0.9212})(10^{1.8142})}{10^{2.5752}}$$
$$= 10^{0.9212+1.8142-2.5752}$$
$$= 10^{0.1602} = 1.45, \text{ approximately.}$$

Needless to say the accuracy obtainable by computation with logarithms is limited by the number of decimal places in the table. But tables are in existence which are carried out to as many places as are needed for all practical purposes.

Remark. Notice that the formula $\log(x + y) = \log x + \log y$, which appears plausible at first glance, is *false*, for $\log x + \log y$ is really $\log(xy)$. If it *were* true we would have an excessively simple table of logarithms. For $\log 1 = 0$; hence $\log 2 = \log(1 + 1) = \log 1 + \log 1 = 0$; and $\log 3 = \log(2 + 1) = \log 2 + \log 1 = 0$, and so on; thus the logarithm of every whole number would be zero and we would need no table of logarithms. If this false formula seemed right to you, it was probably because of its superficial resemblance to the distributive law $g(x + y) = gx + gy$. But the resemblance is no more than superficial. For in the distributive law we are *multiplying* $(x + y)$ by a *number* g, while in the other case we are surely *not* multiplying by the *word* log.

EXERCISES

Evaluate by means of logarithms:

1. $\dfrac{(34.2)(1.57)}{31.3}$.
2. $\sqrt[3]{45.2}$.
3. $\sqrt[5]{117}$.
4. $112(1.04)^{16}$.
5. $\dfrac{21.3}{27.2}$.
6. $\dfrac{(627)(32.4)}{(3.26)(65.8)}$.
7. $\sqrt[3]{67300}$.
8. $\sqrt[5]{96300}$.
9. $376(1.06)^{21}$.
10. $\dfrac{(87.6)(47.4)}{53.7}$.
11. $\sqrt[3]{.00857}$.
12. $\sqrt[5]{.0468}$.

56. Applications of logarithms. An interesting application of logarithms can be made to the compound interest formula. If P dollars are invested at 4 per cent interest compounded annually, the amount at the end of the first year is $P + .04\,P$, or $P(1 + .04)$, or $P(1.04)$. At the end of the second year the amount is the new principal $P(1.04)$ plus the interest on it, which is $.04\,P(1.04)$; thus the amount at the end of the second year is $P(1.04) + .04\,P(1.04) = P(1.04)(1 + .04) = P(1.04)^2$. At the end of n years we would have the amount $A = P(1.04)^n$.

Example 1. \$100 is invested as a trust fund for a child at 4% interest compounded annually. How much will it amount to in 21 years?

$$\begin{aligned} A &= 100(1.04)^{21}. \\ \text{Therefore, } \log A &= \log 100 + 21 \log 1.04 \\ &= 2 + 21(0.0170) = 2.3570. \\ \text{Hence, } \quad A &= \$228, \text{ approximately.} \end{aligned}$$

Example 2. If \$100 is invested at 4% interest compounded quarterly, how much will it amount to in 21 years?

Here the interest will be taken 84 times at the rate of 1% each time. Hence

$$\begin{aligned} A &= 100(1.01)^{84}. \\ \text{Therefore, } \log A &= \log 100 + 84 \log 1.01 \\ &= 2 + 84(0.0043) = 2.3612. \\ \text{Hence, } \quad A &= \$230, \text{ approximately.} \end{aligned}$$

If you think of how long it would take to do this by simple arithmetic, multiplying 100 by 1.01 eighty-four times, the time-saving property of logarithms should now be obvious. Other simple applications will be found in the exercises.

The human race has benefited greatly by saving untold hours because of the existence of tables of logarithms. It is said that when Briggs visited Napier, he remarked,* "My lord, I have undertaken this long journey purposely to see your person, and to know by what engine of wit or ingenuity you came first to think of this most excellent help in astronomy, viz. the logarithms; but, my lord, being by you found out, I wonder nobody found it out before, when now known it is so easy." It is worth noting that this remark could be made with equal justice concerning almost any great mathematical invention. For the highways of mathematics are comparatively easy to follow, since each step is merely one of logical reasoning; but the blazing of the original trail may well have demanded the insight and foresight of genius.

Logarithms are of great value in advanced mathematics quite apart from computation.

EXERCISES

1. If \$124 is deposited in a bank at 4% interest, compounded semi-annually, what will the amount be in 10 years?

2. It is desired to have a sum of \$1000 in the bank ten years from now. The bank pays 4% interest compounded annually. How much should we deposit now?

3. If the \$24 which the Indians received in 1626 for Manhattan had been deposited in a bank paying 4% interest compounded annually, what would it have amounted to in 1940?

4. If a person bets two cents on April 1st and doubles his bet each day thereafter, how much will his bet be on April 30th?

5. The area of the surface of a sphere is $4\pi r^2$ where r is the radius. Taking $\pi = 3.14$ and assuming that the earth is a sphere of radius 3960 miles, find the area of the earth's surface.

6. The volume of a sphere is $\frac{4}{3}\pi r^3$ where r is the radius. Using the data of exercise 5, what is the volume of the earth?

7. If P dollars are invested in a bank at 4% interest, compounded annually, how long will it take for the amount to be $2P$ (that is, doubled)?

8. If \$352 is deposited in a bank at 4% interest, compounded quarterly, what will the amount be in 5 years?

9. The "period" T measured in seconds of a simple pendulum (that is, the time required for a complete oscillation) of length k is given by the formula $T = 2\pi\sqrt{k/g}$. If $k = 3.26$ ft., $g = 32.2$, $\pi = 3.14$, find T.

* Cajori, *History of Mathematics*, 2nd edition, pp. 150–151.

10. Find the length of a pendulum whose period is 1 second, using the formula and data of exercise 9.

11. Suppose that an automobile costing $1000 depreciates at the rate of 20% per year; that is, its value at the end of each year is 80% of its value at the beginning of that year. Find its value at the end of (*a*) 5 years; (*b*) 10 years.

12. Neglecting friction, the velocity, measured in feet per second, of a sled sliding down a hill is given by $v = \sqrt{2\ gh}$ where h is the vertical height of the hill, measured in feet, and $g = 32.2$. If $h = 146$, find v.

57. Calculating machines. The slide rule. Nothing could be further from the truth than the popular notion that mathematicians as a class enjoy doing calculations. The fact is that the mathematicians have always been the ones who disliked the routine performance of tedious calculations sufficiently to do something about it. We have already seen the simplifications introduced into mathematical calculations by the arabic numerals and by the invention of logarithms. We shall discuss here two important mechanical devices designed to further relieve the mind from the necessity of thinking about boring calculations, thereby freeing it to think about more interesting things. The first device is based essentially on the idea of "carrying" in positional notation (arabic numerals) while the second is based essentially on the theory of logarithms.

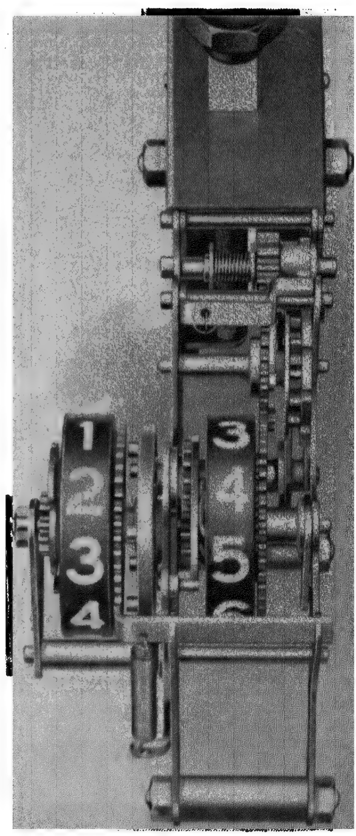


FIG. 40

Calculating machines. The first calculating machines were invented by Pascal and Leibniz in the 17th century. The general idea may be explained as follows, without going into mechanical details. A number of wheels are placed next to each other (Fig. 40). On each wheel appear the numbers 0, 1, 2,···, 9. One may imagine a decimal point to be placed between any two of the wheels or at either extreme. A mechanism is arranged so that a

number may be registered on the wheels by pressing buttons or some other device; that is, this number is made to appear in a horizontal slit in the machine, say at the top of the wheels. Suppose that the number 13075 has been recorded on the machine, and that we now wish to add to this the number 67. Pressing the 7 in the "units column" will cause the "units wheel" to rotate through 7 more spaces thus showing the number 2. The essential point of the machine is that as the units wheel passes from the "nine" to the "zero" it automatically moves the next wheel one space, thus "carrying" one into the "tens column." The machine now reads 13082. Now pressing the 6 in the "tens column," the "tens wheel" rotates through 6 spaces, thus showing the number 4. But as it passes from the "nine" to the "zero," it automatically moves the next wheel one space, thus "carrying" one into the hundreds column. The machine now reads 13142.

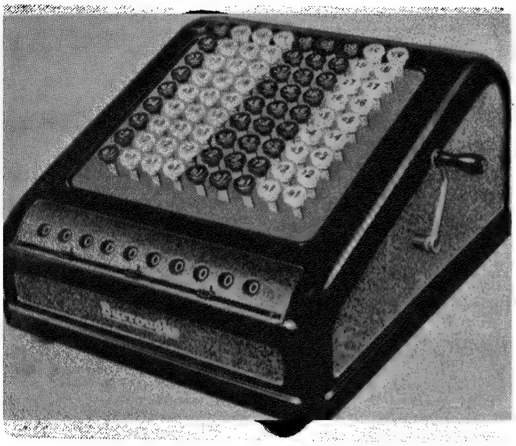


Fig. 41

Multiplication is done essentially by repeated additions. Subtraction and division can also be accomplished on such machines. Needless to say, the actual mechanisms of these machines may be exceedingly ingenious, but at the bottom of the whole thing lies the principle of "carrying" in positional notation. The idea of carrying is also used in the mileage-recorder on an automobile. The actual working of a calculating machine is best understood by using one and examining its movements directly. More detailed discussions of calculating machines may be found in the references at the end of the chapter. Many other kinds of calculating machines have been devised to perform various mathematical operations. Calculating machines, in general, not only relieve the mind of the burden of performing fatiguing and uninteresting calculations but also have the additional virtue of being less likely to make a mistake than a human being. Thus it some-

times is advantageous not to think. But it should not be forgotten that only by thinking can one invent, repair, reconstruct, or even understand the machine.

The slide rule. The rectilinear slide rule was invented by Oughtred, an Englishman, in the 17th century. On each of two adjacent straight bars C and D we mark numbers so that their

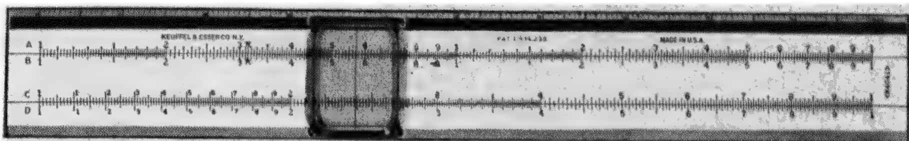


Fig. 42

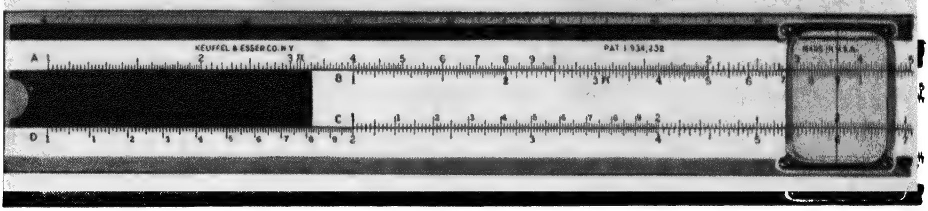


Fig. 43

distances from the 1 mark are proportional to the logarithms of the numbers (Fig. 42). Since log 2 = .301, log 3 = .477 and log 6 = .778, approximately, we might divide the length of each bar into 1000 equal parts and place the number 2 at the 301st mark, the number 3 at the 477th and the number 6 at the 778th. The distance of each mark from the end of the bar is proportional to its logarithm. To multiply 2 by 3 we may proceed to slide bar C so that its beginning (1) is directly over the number 2 on bar D. Then we find the product of 2 by 3 on bar D directly under the number 3 on bar C (Fig. 43). Obviously, sliding the bars in this way has the effect of adding the distance between 1 and 2 on bar D to the distance between 1 and 3 on bar C. But this means essentially adding log 2 and log 3. The resulting distance (between 1 and 6 on bar D) must therefore be log 6 since log 2 + log 3 = log 6.

Similarly to divide 6 by 3 we would place the 3 on bar C over the 6 on bar D and read the result on bar D directly under the 1 on bar C. Clearly we have used the fact that log 6 − log 3 = log 2. Many other operations can be performed with a slide rule.

Further details will best be understood by actually working with one. Since one's eyesight enters into the reading of a slide rule, its accuracy is limited. But it is very useful when great accuracy is not needed, for example, in checking calculations. It is used extensively by engineers and others.

REFERENCES

T. Dantzig, *Number, The Language of Science*, Macmillan, N. Y., 1939.

A. Dresden, *Invitation to Mathematics*, Holt, N. Y., 1936.

H. B. Fine, *College Algebra*, Ginn, Boston, 1904. (Among other things, explanation of the process of extraction of square roots of numbers.)

F. Klein, *Elementary Mathematics from an Advanced Standpoint*, vol. I, Macmillan, N. Y., 1932, (translated by E. R. Hedrick and C. A. Noble.) (*Calculating Machines* pp. 17–22.)

D. E. Smith, *A Source Book in Mathematics*, McGraw Hill, N. Y., 1929. (Contains original accounts by Pascal and Leibniz of their calculating machines and by Oughtred of his slide rule.)

J. E. Thompson, *A Manual of the Slide Rule*, Chapman, London, 1931.

J. W. Young, *Fundamental Concepts of Algebra and Geometry*, Macmillan, N. Y., 1911. (Appendix on "The Growth of Algebraic Symbolism" by U. G. Mitchell.)

Histories of Mathematics, as before.

Chapter VIII

IMPOSSIBILITIES AND UNSOLVED PROBLEMS

58. Introduction. In sections 24 and 38 we have carefully pointed out that it is one thing to say that a problem has not been solved yet and another thing to say that it is impossible to solve it. For example, we proved that it is impossible for any rational number to satisfy the equation $x^2 = 2$; we did not say merely that no one has yet found such a rational number. The distinction between these two things can hardly be called subtle. It must be regarded as one of the world's wonders that many people blithely assume that all professional mathematicians are incapable of grasping this distinction. Every few years, another crank looses a blast against the professional mathematicians' "smug assertion" that angle-trisection, for example, is impossible "merely because they haven't been able to do it." Many of these amazing people go further and claim actually to do it, and their proposed solutions are solemnly published, sometimes privately and sometimes in the daily press. Moreover, these people are difficult to convince. Pointing to the logical proof that angle-trisection, say, is impossible, does no good because if they knew enough mathematics to understand the proof they wouldn't be publishing methods for trisecting angles. On the contrary, they refuse to try to learn the proof of impossibility and insist that the professional mathematician (if they have been so fleet of foot as to corner one) show them exactly where their method is wrong. Since their method usually consists of complicated constructions covering several square yards of paper, this task is not relished by the professional mathematician. Often the misguided zealot has completely misunderstood the problem; it may also happen that he does not grasp how it is possible to prove a thing impossible. The reader has already completed such a proof (section 24).

We shall discuss sketchily some ancient and some modern

problems; some are proved impossibilities and others are merely unsolved or open questions at the present time. Among the classical problems are angle-trisection, the duplication of the cube, and squaring the circle, which are known as the "problems of antiquity." Since mathematics is a living science, there are always a prodigious number of unsolved (that is, open) problems. However, most of those which are really of importance are far too technical to be discussed here. But there are a few unsolved problems, of little practical importance, which are simple enough to be stated here and which are of interest either because of the stubbornness with which they have resisted solution or because of their unconventional nature. Some of these will be discussed below.

59. Constructions with ruler and compasses. The ancient Greeks had a predilection for constructing figures with the aid of straight edge and compasses alone. To begin with there is no reason why we should not equally well restrict ourselves to constructions with ruler alone, or compasses alone, or, on the other hand, why we should not permit other mechanical devices. The problem of constructions under these various conditions has been studied in modern times. The Greeks themselves did many construction problems by means of other devices besides ruler and compasses. It may be conjectured that their preference for ruler and compasses is related to their aesthetic appreciation of circles and straight lines which are, perhaps, the simplest figures. Whether this conjecture is correct or not, the construction problems of antiquity, mentioned above, were supposed to be done with ruler (unmarked straight edge) and compasses alone. For roughly 2000 years these problems were considered open questions and remained a thorn in the side of mathematicians because with the synthetic (purely geometric) methods of the Greeks, no way of settling them was found. Only with the application of algebraic methods to geometry were these constructions proved impossible. These methods were not introduced until modern times and will be discussed in the next chapter.

We shall indicate how one can prove the following (loosely stated) criterion for constructibility with ruler and compasses.

THEOREM. *We can construct a quantity (length) with ruler and*

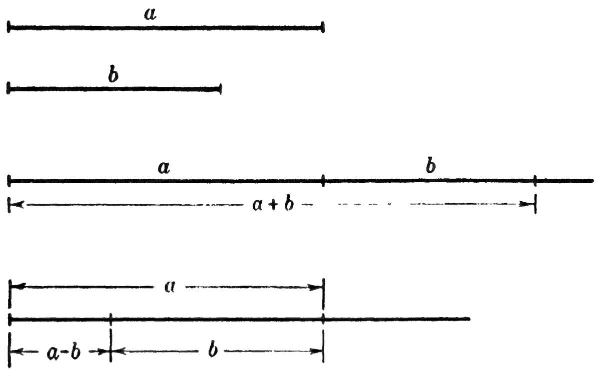


FIG. 44

compasses alone if and only if it can be derived from the data (given lengths) by a finite number of rational operations and extractions of square roots.

This is really a theorem and its converse together (see Remark 2, section 5). The converse (the "only if" part of it) will be discussed at the end of the next chapter. The "if" part of it is easy. We have only to show that given any two lengths a and b (and the unit length) we can construct with ruler and compasses the lengths $a + b$, $a - b$, $a \cdot b$, a/b, and $\sqrt{a}$. The constructions of $a + b$ and $a - b$ are obvious (Fig. 44).

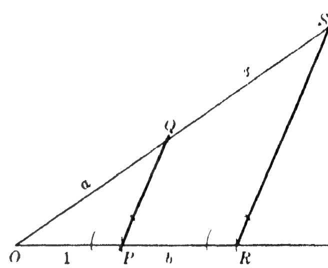


FIG. 45

To construct $a \cdot b$ we lay off a, the unit length, and b on the sides of an angle as in Fig. 45, join P and Q and construct RS parallel to PQ at R. Then $\dfrac{x}{a} = \dfrac{b}{1}$ since parallel lines cut off proportional segments on two transversals. Hence the length $x = a \cdot b$.

To construct a/b we lay off b, the unit length, and a on the sides of an angle as in Fig. 46, join K and L and construct MN parallel to KL at N. Then $\dfrac{y}{1} = \dfrac{a}{b}$ for the reason quoted above. Hence $y = a/b$.

To construct $\sqrt{a}$ we lay off the unit length and a as in Fig. 47, bisect BC, draw a semicircle with the midpoint D as center, and

construct a perpendicular at E. Now angle BFC is a right angle since it is inscribed in a semicircle. Angles BFE and FCE are equal since they are both complementary to angle EFC. Hence the right triangles BFE and FEC are similar because two angles of one are equal respectively to two angles of the other. Hence $\frac{a}{z} = \frac{z}{1}$. Thus $z^2 = a$ or $z = \sqrt{a}$.

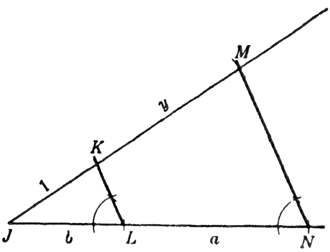


Fig. 46

Now let us consider the problems of antiquity.

60. The duplication of the cube. This problem is known as the Delian problem. It is said that during a plague in Greece, the people went to the oracle at Delos to ask how they might appease the Gods. With the cruel sense of humor often attributed to oracles, it answered that if they would construct a cubical altar with exactly twice the volume of a given cubical altar the plague would stop. The plague ran its course. The people, in their first enthusiasm, are said to have constructed a cube by doubling the side of the given one; but this of course multiplied the volume by 8. Taking the given cube as the unit of volume 1, what they had to do was construct a quantity x to be the side of the doubled cube. But then $x^3 = 2$. The real root of $x^3 = 2$ is $\sqrt[3]{2}$ which can be proved to be not expressible in terms of rational operations and the extraction of square roots only, and hence, by our criterion, is not constructible with ruler and compasses alone.

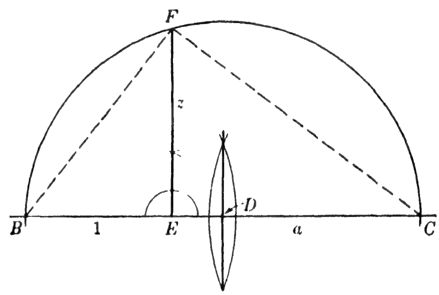


Fig. 47

61. The trisection of a given angle. The problem is to provide a construction by ruler and compasses alone for the trisection of *any* given angle. To be sure, *some* angles, like 180° or 90° can be trisected with ruler and compasses alone; that is, an angle of 60° or 30° can be so constructed. But 120° cannot be trisected; that is, an angle of 40° can not be constructed, with ruler and compasses alone. This can be established by our criterion. It will

not be done here. For practical work, anybody can trisect any angle approximately enough by using a protractor finely enough divided. Theoretically exact solutions can also be obtained by mechanical devices other than ruler and compasses alone. Neither of these solutions have to do with the classical problem. Angle-trisection and circle-squaring are still popular sports among the vast army of amateur cranks; a proposed "solution" for one or the other appears every few years.

62. Squaring the circle. The problem is to construct a square whose area is exactly equal to that of a given circle. Let the radius of the circle be 1. Then the area of the circle is π. But if x is the side of the square to be constructed we must have $x^2 = \pi$. If we could construct a length equal to π we could also construct $x = \sqrt{\pi}$, and conversely. But it can be shown that x is not constructible because π is not constructible by ruler and compasses alone. In fact, not only is π inexpressible in a finite number of rational operations and square roots, but it is even inexpressible in terms of roots of any index. This is a consequence of the theorem that π does not satisfy any polynomial equation with integral coëfficients, proved by Lindemann in 1882. There are, however, mechanical devices which can construct π.

A number which satisfies a polynomial equation with integral coefficients is called an **algebraic number.** A number satisfying no such equation is called a **transcendental number.** Hence π is a transcendental number. A rational number $x = p/q$, where p and q are integers, by definition; hence x satisfies the equation $qx - p = 0$ with integral coefficients. It follows that every rational number is algebraic, and hence that every transcendental number is irrational. On the other hand $\sqrt{2}$ is algebraic, although irrational, since it satisfies the equation $x^2 - 2 = 0$ with integral coefficients. Whether the number π^π is algebraic or transcendental is still unknown today.

63. Construction of regular polygons. A polygon is called **regular** if its sides and angles are all equal. By using our criterion it can be shown that while regular polygons of 4, 5, 6, 10 sides and others can be constructed, a regular polygon of 7 sides cannot. It was proved by Gauss, at the age of 19, that a regular polygon of 17 sides can be constructed, a fact which was not

previously suspected. He also gave such a construction. In fact, Gauss proved that if the number $X = 2^{2^n} + 1$ is prime then a regular polygon of X sides can be constructed. A **prime number** is a natural number, greater than one, which has no factors except 1 and itself, like 2, 3, 5, 7, 11, 13, 17, 19, $\cdots$. For $n = 1$ and $n = 2$ respectively, $X = 5$ and $X = 17$. For $n = 3$, $X = 257$, a prime number. Detailed discussions of the construction of the regular polygon of 257 sides have been given. For $n = 4$, $X = 65{,}537$ is also prime. The regular polygon of 65,537 sides has been discussed. If a prime number X is not expressible as $2^{2^n} + 1$ then a regular polygon of X sides cannot be constructed with ruler and compasses alone.

Leonhard Euler
1707–1783, Swiss

64. Problems about prime numbers. The preceding section suggests that it might be desirable to know for what natural numbers n the number $2^{2^n} + 1$ is prime. Fermat (see section 65) believed that $2^{2^n} + 1$ was prime for all natural numbers n. Euler (1707–1783) showed that this is not so by finding that 641 is a factor of $2^{2^5} + 1 = 4{,}294{,}967{,}297$. However, there remains the question of whether the unending sequence of numbers $2^{2^n} + 1$ as $n = 1, 2, 3, 4, 5, \cdots$ contains only a finite (limited) number of primes or an infinite number of primes. This question is still an open or unsolved problem.

There are many unsolved problems connected with prime numbers such as the following:

1. Is every even number, other than two, the sum of two primes? (Goldbach's conjecture.) For example, $4 = 2 + 2$, $6 = 3 + 3$, $8 = 5 + 3$, $10 = 7 + 3$, $12 = 7 + 5$. No one has

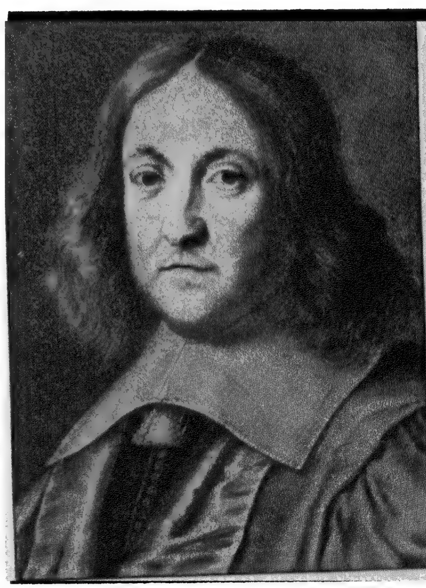
ever found an even number greater than 2 which is not so expressible, but neither has anyone been able to *prove* that they all can be so expressed.

2. Find a formula for the number of primes between two given natural numbers.

3. Is there an infinite number of pairs of primes differing by two? For example, 11 and 13, 17 and 19, 29 and 31.

4. How can one decide practically whether or not a given natural number is prime? Clearly one can decide whether or not any natural number is prime by actually trying to divide it by every smaller natural number.* But this is not practical for large numbers. A machine was invented by D. H. Lehmer a few years ago which can decide question 4 in a short time for many very large numbers. For example, it factored $2^{93} + 1$ into 529,510,939 times 2,903,110,321 times 715,827,883 times 3^2 in 3 seconds.†

65. Fermat's last theorem. Pierre de Fermat (1601–1665) was a lawyer and councillor for the parliament of Toulouse, whose leisure time was given largely to mathematics. He left his mark on many branches of mathematics, but had the annoying habit of making brief marginal notes of his discoveries. One of these marginal notes, discovered after his death, asserted the following theorem: *if n is a natural number greater than 2 there cannot be three natural numbers x, y, and z, such that $x^n + y^n = z^n$.* (Notice that for $n = 2$ there *are* such numbers; for example $3^2 + 4^2 = 5^2$.) He wrote



Pierre de Fermat
1601–1665, French

* This method can be somewhat refined to diminish the labor a little but it remains impractical for very large numbers.

† Ball, *Mathematical Recreations and Essays* (11th edition), p. 61. For a description of the machine see D. H. Lehmer, "A Photo-Electric Number Sieve," *American Mathematical Monthly*, 40, pp. 401–406 (1933).

that he had a remarkable proof for this for which there was no room in the margin. Whether or not Fermat actually had a proof remains exceedingly doubtful because no one has yet proved the theorem, although many erroneous proofs have been proposed in the intervening three centuries. The theorem *has* been proved for *certain* values of n including every value from 3 up to 616.* Note also that if the theorem has been proved for any value of n it is automatically proved for all multiples of that value. For example, having proved the theorem for $n = 3$ it is automatically proved for $n = 6$. For if there *were* integers x, y, z satisfying $x^6 + y^6 = z^6$ then the integers x^2, y^2, z^2 would satisfy the relation $(x^2)^3 + (y^2)^3 = (z^2)^3$. But no one has ever been able to prove it for *all* values of n. It is interesting to note that while the problem itself is not of great practical importance, the attempts to solve it have been the source of many important mathematical researches. Large prizes, still unclaimed, have been offered for the solution of this problem. It remains an open question.

66. The four-color problem. The following problem might have originated in a printing shop which made colored maps.

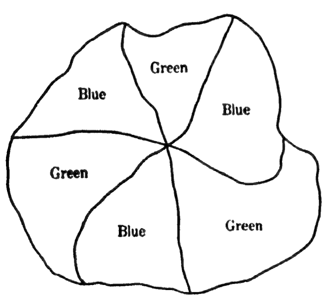


FIG. 48

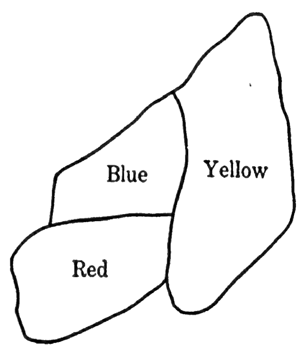


FIG. 49

Suppose we agree to color every map according to the following rules:

1. If two countries have a strip of boundary in common they *must* have different colors;

2. If two countries have only isolated points of boundary in common they *may* have the same color.

* See H. S. Vandiver, *Duke Mathematical Journal* III (1937) pp. 569–584.

Each country is understood to be in one connected piece. For example, the coloring of Fig. 48 is permissible. Notice that the map in Fig. 48 cannot be colored with less than 2 colors. The map in Fig. 49 *requires* 3 colors. The map in Fig. 50 *requires* 4 colors. The problem is: does there exist a map which *requires* 5 colors? Or to put it another way: *can* every map be colored with only 4 colors? This problem can be traced back to about 1840. Many people have proposed proofs that every map can be colored with 4 colors; all these proofs have been erroneous. It *has* been proved that every map can be colored with 5 colors, but no one has ever produced a map that really *required* 5 colors; all known maps, at the present writing, *can* be colored with only 4 colors. Whether the number is 4 or 5 remains an open question today. Curious partial results have been proven. For example, it has been proved recently that if there were a map requiring 5 colors it would have to have at least 36 countries on it. The problem has an extensive literature.

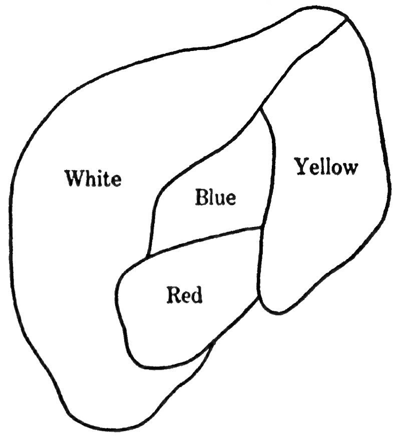


FIG. 50

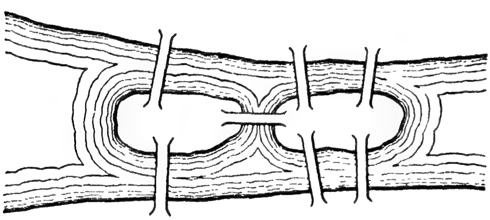


FIG. 51

67. The seven bridges of Königsberg. The city of Königsberg, in the 18th century, had seven bridges situated as in Fig. 51. The problem is: can a pedestrian cross all seven bridges without crossing any one of them twice? The pedestrian may start anywhere and finish anywhere. This problem was solved in 1736 by Euler, who proved it impossible. The problem may be studied as follows: choose one point or vertex on each island and on each shore and join these vertices by paths across the bridges. We get thus the figure in Fig. 52, consisting of 4 points

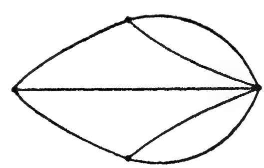


FIG. 52

joined by 7 paths. The question becomes: can we trace these paths with a pencil without lifting the pencil from the paper and without tracing any path twice. Many problems of this nature appear in our newspapers. They may be settled as follows. A vertex is called **even** or **odd** according as it has an even or odd number of paths emanating from it. Then it is not hard to prove that a figure can be traversed exactly once, as required above, if and only if the number of odd vertices is either 0 or 2. In fact, if the number of odd vertices is zero, then we may start at any vertex and end at the same place; if the number of odd vertices is 2, we may start at one of them and end at the other. Thus the figure in Fig. 53 can be traversed exactly once beginning at the odd vertex A and ending at the odd vertex B. Try to see why this is so. The bridge problem above is an impossibility, and not an open question, since there are 4 odd vertices. Further details may be found in Ball, *Mathematical Recreations and Essays.*

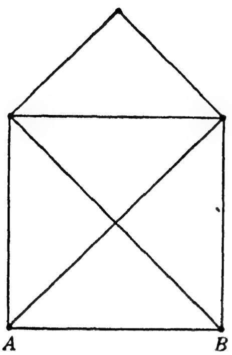


FIG. 53

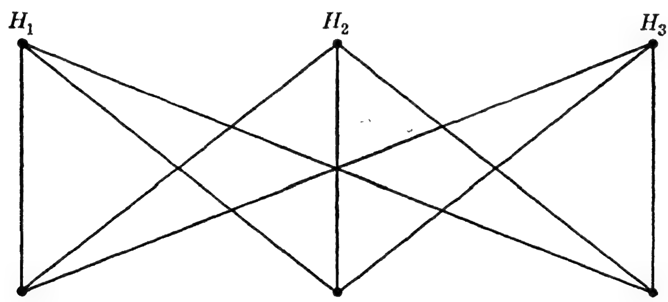
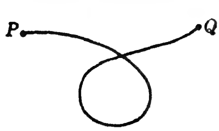


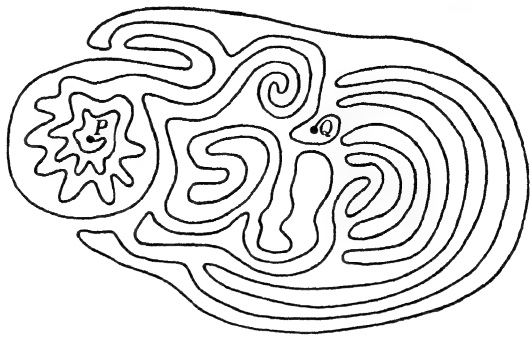
FIG. 54

68. Three houses and three wells. The following is a very familiar puzzle. There are three houses, H_1, H_2, H_3 and three wells W_1, W_2, W_3 (Fig. 54). We are to join each house to each well by a pipe but no pipe is permitted to cross any other pipe. Translated into mathematical language, we are to join each of the vertices H_1, H_2, H_3 to each of the vertices W_1, W_2, W_3 by a simple path, no two paths



Not a simple path
FIG. 55

being permitted to cross. (A **simple** path is one which does not cross itself. The path joining P and Q in Fig. 55 is not simple, while that of Fig. 56 is simple.) Hence the joining in Fig. 54 is *not* permissible. Experiments will indicate (Fig 57) that we can insert 8 of the 9 paths without crossings, but not the 9th. (Note that a pipe may not be passed under a house. The houses are mathematically considered to belong to the paths as end-points.) We will sketch below an intuitive proof of the impossibility of this problem. Some preliminaries will be needed.



A simple path

FIG. 56

Consider the figure in Fig. 58, consisting of a simple path in the plane with its two vertices (end-points). By a **network** we shall mean any figure which can be built up from Fig. 58 by means of any succession of the following two operations:

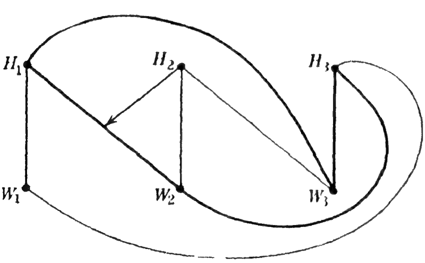


FIG. 57

(1) Adding a new vertex and joining it to an old vertex by a simple path not crossing any existing path, as in Fig. 59;

(2) Joining two existing vertices by a new simple path not crossing any existing path, as in Fig. 60.

Examples of networks are to be seen in Figs. 52, 58, 59, 60, 61.

We have the following theorem due to Euler.

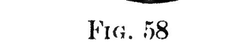


FIG. 58

THEOREM. *Let V be the number of vertices, and E the number of paths in any network situated in a plane. If the network divides the plane into F regions, then $V - E + F = 2$.*

For example, the network in Fig. 58 has $V = 2$, $E = 1$, and $F = 1$ since it "divides" the plane into one region, that is, leaves

the plane undivided; clearly $2 - 1 + 1 = 2$. The network in Fig. 59 has $V = 3$, $E = 2$, $F = 1$; obviously $3 - 2 + 1 = 2$.

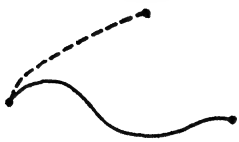


Fig. 59

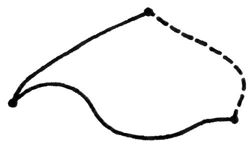


Fig. 60

The network in Fig. 60 has $V = 3$, $E = 3$, and $F = 2$ since it divides the plane into 2 regions, an inside and an outside. In the network of Fig. 61, before inserting the dotted path AB by means of operation (1) and the dotted paths CD and EF by means of operation (2), we have $V = 8$, $E = 10$, $F = 4$, and $8 - 10 + 4 = 2$. After inserting the dotted paths we have $V = 9$, $E = 13$, $F = 6$ and $9 - 13 + 6 = 2$.

To prove the above theorem we have only to reason as follows. The simple network of Fig. 58 clearly yields 2 as the value of the expression $V - E + F$. But every network can be built up from this one by means of operations (1) and (2), by definition. Now operation (1) cannot alter the value of $V - E + F$ since it leaves F unchanged but increases V by one and E by one, and V and E have opposite signs. Similarly, operation (2) leaves the value of V − E + F unaltered since it leaves V unchanged but increases E by 1 and clearly increases F by 1; but E and F have opposite signs. It follows that every network must have the same value for $V - E + F$ as the network of Fig. 58, namely 2. This completes the proof of Euler's theorem.

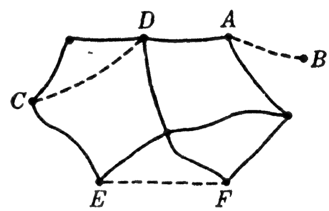


Fig. 61

We are now in a position to establish that the 3 houses and 3 wells puzzle is impossible, by means of an indirect proof (reductio ad absurdum). Suppose it were possible. Then we would have a network in the plane with 6 vertices and 9 paths. That is, $V = 6$, and $E = 9$. From Euler's theorem we have $6 - 9 + F = 2$; hence $F = 5$. Inspection of Fig. 54 reveals that there are no three vertices each pair of which are joined by a path; that is, there are no "triangular" regions. That is, there are no regions bounded by exactly 3 paths. Similarly, inspection reveals that there are no regions bounded by exactly 2 paths; that is, there is no pair of vertices joined by exactly 2 paths. Therefore each

region must be bounded by at least 4 paths. Since there are 5 regions, we might now be tempted to think that there must be at least $4 \cdot 5 = 20$ paths; but this would be wrong because each path is on the boundary of exactly two regions and has therefore been counted twice. It *is* correct, however, to say that there must be at least half of $4 \cdot 5$ paths, or at least 10 paths. This contradicts the hypothesis that our network is to have exactly 9 paths. Since our supposition that the problem is possible leads to a contradiction, it must be impossible.

Exercise. Imitating the proof in the preceding paragraph as far as possible, show that it is impossible to join each of the five vertices of a pentagon to every other vertex by simple paths (in a plane) which do not cross each other. (Hint: if it could be done there would be a network with $V = 5$, $E = 10$. See Fig. 62).

A remarkable and difficult theorem proved by Kuratowski in 1930 asserts that any network which cannot be drawn in a plane, without crossings, must contain in it either the network of Fig. 54 or that of Fig. 62.

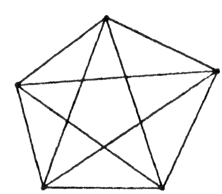
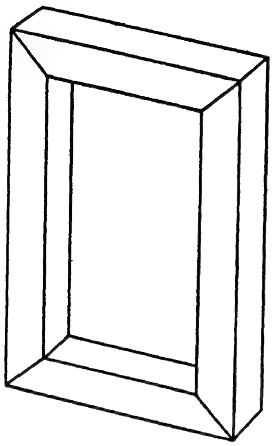


Fig. 62

While *our* interest in networks has been due to a mere puzzle, it is worth remarking that networks were studied by Kirchoff to good advantage in connection with his work on electrical circuits (1847).

Remark. Euler's theorem applies equally well to a network drawn on the surface of a sphere. If one thinks of the part of the plane in which our network is drawn as being a square sheet of rubber on which we place a sphere, resting on its south pole, we can stretch the rubber sheet over the surface of the sphere and "button it up" at the north pole, without altering the numbers V, E, F. Now imagine an ordinary polyhedron (like a cube, or a pyramid) made of rubber and placed within a glass sphere. The rubber polyhedron can be blown up until it clings to the inner surface of the glass sphere. Then the vertices, edges, and faces of the polyhedron appear as vertices, paths, and regions of a network on the sphere. Hence, for any polyhedron which can be so deformed into a sphere we have $V - E + F = 2$ where V is the number of vertices, E the number of edges, and F the number of faces. For example, a cube has 8 vertices, 12 edges, and 6 faces, and $8 - 12 + 6 = 2$. Euler's theorem is *not* true for networks

drawn on the surface of a doughnut or inner tube, for example. Therefore, the corresponding theorem for polyhedra like that of a picture frame (Fig. 63), which could be placed within an inner tube and blown up as above, is no longer true. For such networks or polyhedra we have $V - E + F = 0$. Other surfaces yield different values for the number $V - E + F$.



$V=16, E=32, F=16, \quad V-E+F=0$

Fig. 63

69. Puzzles. Many of the problems mentioned in this chapter are of interest only as puzzles rather than because of any practical value. As we pointed out before, this is partly because unsolved problems of practical value would be too technical to discuss here. One should not get the impression that mathematicians concern themselves exclusively with problems like those in this chapter. In fact, a student who brings a puzzle to his teacher is frequently met with rebuff. This is often due to the fact that most puzzles which circulate among laymen either depend on a play of words or joke of some sort, or, if legitimate, are easily soluble by well known mathematical methods. For the mathematician, unlike the average layman, is not content to solve an individual puzzle; he immediately generalizes the problem and tries to find a general method of solution for a whole class of similar problems. For example, Euler did not stop with the solution of the Königsberg bridge problem but found the solution of a whole class of problems of that sort.

While it would be frightfully wrong to regard mathematics as nothing but a collection of tricks for solving puzzles, an unfortunate impression which some students seem to get, nevertheless there is something to be said for the lowly puzzle. There can be no doubt that the personal satisfaction derived from a successful attack on a puzzle is a universal and human emotion. In fact, the somewhat childish "I-can-do-something-you-can't-do" spirit may have stimulated mathematical research considerably.*

* It should not be supposed, however, that mathematicians are *never* actuated by a pure love of science. A. Einstein says in his prologue to M. Planck's *Where Is Science Going?*: "Many kinds of men devote themselves to science, and not all for the sake of science herself. There are some who come to her temple because it offers them the op-

For example, during the 16th and 17th centuries it was customary for a mathematician who had solved a new and difficult problem to withhold his method of solution and challenge all comers to solve the problem. This custom led to many entertainingly undignified and often violent disputes, but also served to stimulate other mathematicians who frequently advanced the progress of mathematics by their efforts. In modern times, however, new results are published for all to read; this seems to be a much more mature way to cooperate for the advancement of science.

It should also be said that even the lowliest and most unimportant looking puzzle may possibly lead to general and important considerations. We have already seen that apparently simple problems have led to mathematics of such difficulty that they are still unsolved. In any case, much ingenuity has been expended on what may be called mathematical amusements or recreations. Hosts of people have found in mathematical puzzles and in mathematics in general a never-failing source of enjoyment. As a hobby mathematics has the very desirable characteristic of being inexhaustible. For even if you were to solve all the problems in existence at the moment (an unlikely eventuality), you could proceed to invent more problems for yourself.

REFERENCES

W. W. R. Ball, *Mathematical Recreations and Essays* (11th edition), Macmillan, London, 1939. (More complete discussions of many of the problems mentioned in this chapter, and many others.)

E. T. Bell, *Men of Mathematics*, Simon and Schuster, N. Y., 1937. (Biographies of Fermat, Euler, Gauss, et. al.)

P. Franklin, "The Four-Color Problem," *Scripta Mathematica*, vol. VI, 1939, p. 149 and p. 197.

E. Kasner and J. Newman, *Mathematics and the Imagination*, Simon & Schuster, N. Y., 1940.

A. de Morgan, *Budget of Paradoxes*, Open Court, Chicago, 1940. (The angle-

portunity to display their particular talents. To this class of men science is a kind of sport in the practice of which they exult, just as an athlete exults in the exercise of his muscular prowess. There is another class of men who come into the temple to make an offering of their brain pulp in the hope of securing a profitable return. . . . Should an angel of God descend and drive from the Temple of Science all those who belong to the categories I have mentioned, I fear the temple would be nearly emptied. But a few worshippers would still remain—some from former times and some from ours."

trisectors, circle-squarers, *et. al.*, are flayed in a delightful manner by a master of academic sarcasm.)

J. W. A. Young, *Monographs on Modern Mathematics*, Longmans Green, N. Y., 1911. (Chapter on Constructions with Ruler and Compasses by L. E. Dickson, and Chapter on the Theory of Numbers by J. W. A. Young.)

Histories of Mathematics, as before.

Chapter IX

ANALYTIC GEOMETRY

70. Introduction. The geometry of the ancient Greeks was the outstanding scientific achievement of the human race up to their time and for hundreds of years thereafter. The long period called the Dark Ages began, so far as mathematics is concerned, roughly with the fall of Greek civilization. The Romans, preoccupied with military conquest, governmental affairs, and the acquisition of riches, contributed nothing to the progress of science. In 529 A.D., the emperor Justinian expressed his disapproval of heathen learning by closing the schools at Athens. Europe, for roughly a thousand years, preserved in its monasteries barely enough science, and that often copied rather than understood, to calculate the calendar in order to ascertain the dates of religious holidays. Not only were no improvements made in Mathematics but the excellent work of the Greeks was forgotten or degraded throughout Europe. Only the Arabs kept alive an interest in science and they preserved much of the Greek work together with Hindu arithmetic during these centuries. With the beginning of the Renaissance, progress in Mathematics revived along with progress in other fields. The algebra we have studied was developed largely from the 13th century to the first years of the 17th century, its slow advance suffering long interruptions due to frequent periods of war and religious strife.

Yet, almost miraculously, the thread of science has persisted through the centuries, often disappearing almost entirely. To be sure, much ancient work has been lost. A considerable number of Greek works were destroyed, for example, in the tragic burning of the great library at Alexandria, containing hundreds of thousands of books collected for many centuries, in the 7th century A.D. by the Mohammedan conquerors. Their leader is said to have held that if the books disagreed with the Koran they lied and should be destroyed whereas if they agreed with the

Koran they were superfluous. The books were therefore burned to heat water for the soldiers' baths. This was more serious than similar modern instances of religious, racial, and nationalistic intolerance since it occurred before the invention of the printing press, so that few, if any, copies of any given work could be expected to exist. The early Mohammedans were not the first to damage the treasures of the Alexandrian library; they were merely the last and possibly the most thorough. The library had been subjected to vandalism before by other zealots. Fortunately, copies of some Greek masterpieces remained unharmed, sometimes in remote places. One of Archimedes' works, for example, was found at Constantinople as recently as 1906 by Heiberg.

The 15th and 16th centuries saw the invention of the printing press, the circumnavigation of the earth, and the beginning of the struggle against dogmatism and repressive authority. The Renaissance, in all its phases, got fairly under way. With the 17th century, there began a period of tremendous mathematical activity which is still in progress today. A great deal of modern mathematics has its roots in the 17th century and came to flower in the 18th and 19th centuries. The feats of engineering, the wonders of mathematical physics and astronomy, the manifold applications of mathematics to the field of statistics, all of which are so commonplace today, were made possible by mathematical researches which began in the 17th century.

Until the 17th century, geometry had been thought of as dealing with figures in space and algebra as dealing with numbers, and no one had thought of studying geometry systematically by means of algebra. In the first half of the 17th century, however, several people, notably Descartes and Fermat, saw that geometric theorems could be interpreted algebraically, and that algebraic theorems could be interpreted geometrically. In 1637, Descartes published a famous work on geometry in which he studied with great ease, by his algebraic methods, geometrical problems which demanded the greatest ingenuity when attacked solely with the purely geometrical methods of the Greeks. This so-called *analytic geometry*, *coordinate geometry*, or *Cartesian geometry* (from Cartesius, the latinized form of Descartes) was part of the foundation on which the calculus of Newton and

Leibniz was reared in the latter half of the 17th century. In fact, it may be said to mark a turning point in the development of mathematics. Its importance lies in the fact that it unifies algebra and geometry. Let us see how this union of algebra and geometry was accomplished.

Since elementary algebra is principally concerned with numbers (or expressions involving numbers) and geometry deals chiefly with points (or figures composed of points), it is natural to expect the union of the two subjects to be achieved by somehow associating points with numbers. This is done essentially by attaching to each point at least one label, each label consisting of one or more numbers. The numbers composing a label attached to a given point are called **coordinates** of the point. When we have attached such numerical coordinates to all our points in any way whatever we say that we have set up a **coordinate system** in our geometry. Such an association or correspondence between numerical labels and points may be made in any manner whatever; but if it is to be useful, a coordinate system should have certain regular properties. For example, if P and Q are different points, no label attached to P should be exactly the same as any label attached to Q. Also, if two points are near to each other, they should have coordinates which do not differ very much. In setting up such a coordinate system we can hardly meander about tying labels to points in a haphazard way one point at a time. Hence we use some device or scheme which enables us to attach coordinates to all points in a systematic way, so that we can determine the coordinates of any given point or the point having any given coordinates easily. Hence we may identify any point by means of its numerical coordinates or label. Some of the usual schemes for doing this will now be taken up.

71. One-dimensional geometry. Consider a single straight line. We have already seen how we can associate a real number with every point on it, after having chosen a starting point or origin, a unit of length, and a positive direction. It is customary

the right side of the line.* The real point is called its **coordinate.** To ›orresponds a definite real number

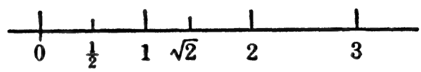


FIG. 64

re corresponds a definite point on *respondence* between real numbers led a **one-dimensional coordinate** use each point is associated with a sen a different point as origin, or a fferent positive direction, we would ·dimensional coordinate system for

Let us choose a definite one-dimensional coordinate system on our line. The distance from the origin to the point whose coordinate is 3 is 3. The distance from the origin to the point whose coordinate is -3 is also 3, since distance is always understood to be a positive number. By the **absolute value** of the real number x we mean x itself if x is not negative, or the positive number $-x$ if x is negative. We denote the absolute value of x by $|x|$. For example, $|3| = 3$ and $|-3| = 3$. **The distance from the origin of the point whose coordinate is x is $|x|$.**

The distance between the points whose coordinates are 1 and 3 respectively is $|1 - 3| = |-2| = 2$. The distance between the points whose coordinates are -1 and -3 respectively is $|(-1)-(-3)| = |2| = 2$. The distance between the points whose coordinates are 1 and -3 respectively is $|1 -(-3)| = |4| = 4$. We could prove the following theorem.

THEOREM 1. *The distance between two points whose coordinates are x_1 and x_2, respectively, is $|x_1 - x_2|$.*

* If you turn the page upside down, or stand on your head, the "right" side becomes the "left." This cannot affect the geometry of the figures we study since the figures are surely unconcerned about our posture.

This theorem may also be stated as follows: *the distance between the points whose coordinates are x_1 and x_2 respectively is whichever of the two numbers $x_1 - x_2$ or $x_2 - x_1 = -(x_1 - x_2)$ happens to be positive.* If the point with coordinate x_1 is to the right of the point with coordinate x_2, then $x_1 - x_2$ is positive. *Another way to indicate $|x_1 - x_2|$ symbolically is $\sqrt{(x_1 - x_2)^2}$,* since we have agreed that the radical sign denotes the positive square root, wherever possible. For example

$$|(-3)-(-1)| = \sqrt{[(-3)-(-1)]^2} = \sqrt{[-2]^2} = 2.$$

Let x_1 and x_2 be the coordinates of two different points. Let x' be the coordinate of their midpoint; that is, the point midway between them. Then either $x_1 < x' < x_2$ or $x_1 > x' > x_2$. Suppose

14 N

FIG. 65

the first case holds (Fig. 65). Then $x' - x_1$, being positive, is the distance between those two points, and $x_2 - x'$, being positive, is the distance between these two points. But x' is the midpoint. Hence we must have

(1) $$x' - x_1 = x_2 - x'.$$

Hence, $$x' + x' = x_1 + x_2$$

or $$2x' = x_1 + x_2.$$

Finally, $$x' = \frac{x_1 + x_2}{2}.$$

Exercise. The other case, where x_1, x', x_2 are in order from right to left, proceeds similarly. Give the proof for this case.

We have the following theorem.

THEOREM 2. *The coordinate x' of the midpoint between the points whose coordinates are x_1 and x_2 is given by $x' = \frac{1}{2}(x_1 + x_2)$; that is, it is the average of the given coordinates.*

For example, the coordinate of the midpoint between the points whose coordinates are 5 and (-3) is $\dfrac{5 + (-3)}{2} = 1$.

EXERCISES

Find (a) the distance between the points whose coordinates are given; (b) the coordinate of their midpoint:

1. 7 and 5. **2.** 6 and -2. **3.** -3 and -7. **4.** 4 and 8.
5. -3 and 7. **6.** -5 and -1. **7.** 8 and 0. **8.** -4 and 0.
9. 5 and 2. **10.** -3 and 8.

11. The point whose coordinate is 2 is the midpoint between the point whose coordinate is -3 and another point. Find the coordinate of the other point.

72. Two-dimensional geometry. We now turn to plane geometry with which we shall be occupied for most of this chapter. We may introduce coordinates into the plane as follows.

Choose two perpendicular lines, and a unit of length. Call one of the lines the **x-axis** and the other the **y-axis**, and call their point of intersection the **origin.** Choose a positive direction on each of the axes and, by means of the chosen unit of length, set up a one-dimensional coordinate system on each axis. It is customary to consider the x-axis horizontal and the y-axis vertical, and their positive directions as being to the right and up respectively,* as in Fig. 66. Now consider any point P in the plane. Draw a line through P perpendicular to the x-axis. The number attached to the point on the x-axis where this perpendicular meets it is called the **x-coordinate** or **abscissa** of P. Similarly, draw a line through P perpendicular to the y-axis; the number attached to the point on the y-axis where this perpendicular meets it is called the **y-coordinate** or **ordinate** of P. Thus the abscissa of the point P in Fig. 66 is 3 and its ordinate is 1; *we indicate these facts briefly by writing that the coordinates of P are* (3,1), *writing the x-coordinate first.* Thus the coordinates of Q in Fig. 66 are $(-1,0)$ the coordinates of R are $(0,-1)$, and the coordinates of the origin are (0,0). Clearly, each point in the

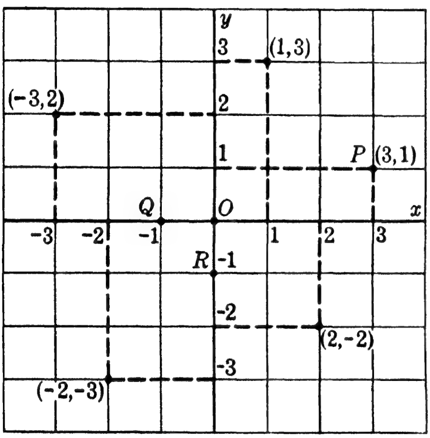


FIG. 66

* See the footnote on page 216.

plane has a definite pair of coordinates and every pair of real numbers determines a definite point in the plane. This *one-to-one correspondence* between the points of the plane and number

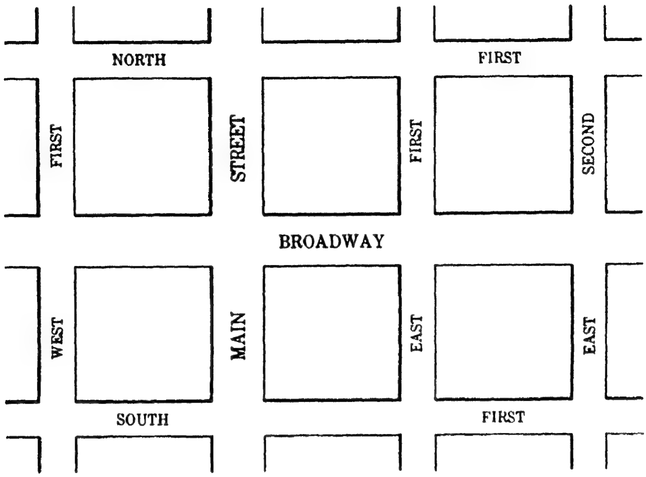


Fig. 67

pairs (x,y) is called a **two-dimensional coordinate system,** *two*-dimensional because each point is identified by a *pair* of real numbers. If we had chosen a different pair of axes, or a different unit of length, or different positive directions, we would have obtained a different two-dimensional coordinate system for the same plane. In most of this chapter we shall suppose that we have chosen a definite two-dimensional coordinate system. We shall speak loosely of the point (x,y) or the point $(2,3)$, meaning the point whose coordinates are (x,y) or $(2,3)$, respectively. Lines parallel to the x-axis are called **horizontal** and lines parallel to the y-axis are called **vertical.** A line which is neither horizontal nor vertical is called **oblique.**

The idea of coordinates introduced here is nothing but a systematization of the way in which we might locate a certain corner in a city with rectangular streets, such as the corner of East 2nd St. and South 1st St. To reach this corner from Main and Broadway, we go 2 blocks East and 1 block South. Similarly, to say that the coordinates of a point are $(2,-1)$ means that to reach it, starting from the origin, we must proceed 2 units in the direction of the positive side of the x-axis and then 1 unit in the direction of the negative side of the y-axis. Marking a point with given coordinates is called **plotting** the point.

In this simple idea of coordinates we have the germ of the unification of geometry and algebra. Although geometry may be concerned with points and algebra with numbers, we have now identified each point in the plane with a pair of numbers and conversely. It remains now to exploit and develop this idea. In particular, we shall see how various geometric concepts can be translated into algebraic terms, by means of this simple device of coordinates, and studied by algebraic means. It may be difficult to see how so simple a notion can lead to anything important. This, however, can be said of many great scientific ideas. It takes insight and imagination to see in a simple idea its implications and potentialities.

Remark. Squared graph paper will be useful for all the exercises in this chapter. A diagram should be made for each exercise.

EXERCISES

1. Plot the points whose coordinates are $(1,2)$, $(2,1)$, $(-1,2)$, $(2,-1)$, $(-2,-1)$.

2. Plot the points whose coordinates are $(3,2)$, $(2,3)$, $(3,-2)$, $(-3,2)$, $(-3,-2)$.

3. Plot the points whose coordinates are $(3,0)$, $(0,3)$, $(-3,0)$, $(0,-3)$, $(0,0)$.

4. What is the y-coordinate of any point on the x-axis?

5. What is the x-coordinate of any point on the y-axis?

6. What are the coordinates of the origin?

7. If two points are on the same vertical line what can be said about their x-coordinates?

8. If two points are on the same horizontal line, what can be said about their y-coordinates?

9. A square has its center at the origin. (The center of a square is the point at which its diagonals intersect.) The sides of the square are parallel to the x- and y-axes respectively. If one of the vertices of the square is the point $(3,3)$, find the coordinates of each of the other three vertices.

10. A rectangle has vertices at $(0,0)$, $(5,0)$, $(0,3)$. Find the coordinates of the fourth vertex.

73. Other types of coordinate systems. The two-dimensional coordinate systems introduced into the plane in the preceding section are known as **rectangular coordinate systems** because the axes are perpendicular. We digress for a moment to point out that other types of coordinates can be used. For example, we

might use what are called **oblique** coordinate axes as in Fig. 68.

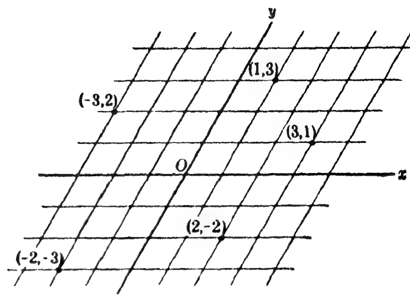


Fig. 68

Or, instead of saying that a certain point may be reached by proceeding 1 mile East and 1 mile North (rectangular coordinates), we might say walk at an angle of 45° between North and East for a distance of $\sqrt{2} = 1.414 \cdots$ miles (Fig. 69). This would be most natural in open country, for example. This suggests that any point in the plane can be specified by stating the angle A (measured in the counterclockwise direction, say) with a chosen fixed direction, and the distance r to be traversed from the origin O, or starting point, in the direction given by the angle (Fig. 70). If we specify our points by such a pair (A, r) we call the resulting correspondence between number-pairs (A,r) and points a **polar coordinate system.**

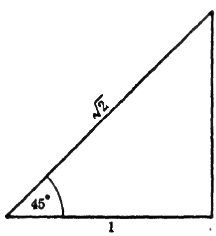


Fig. 69

Other systems, such as various kinds of **curvilinear coordinates** are sometimes used in which systems of curves take the place of the straight lines on ordinary graph paper (Fig. 71). If one were studying the geometry of figures on a curved surface of some sort instead of a plane (flat surface), we might naturally employ some sort of curvilinear coordinates. An example of this is our system of latitude and longitude on the surface of the earth. When we specify a place as 60° west longitude and 42° north latitude we are using the curvilinear coordinate system pictured in Fig. 72 with the equator and Greenwich meridian as x and y axes respectively. The "parallels" of latitude and "meridians"

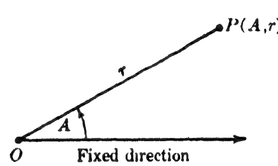


Fig. 70

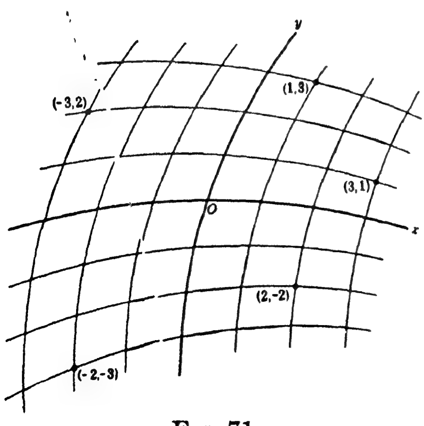


Fig. 71

of longitude form the network of curves. Degrees are used in this system because longitude and latitude signify angles at the center C of the earth as indicated in Fig. 72.

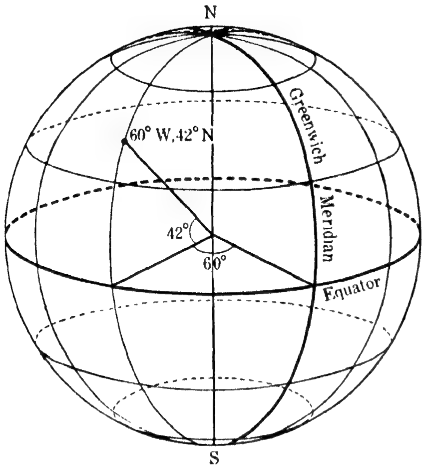


FIG. 72

In plane geometry, the rectangular coordinates introduced in the preceding section are the easiest to use. Hence in our study of plane geometry here, *we shall use rectangular coordinates exclusively.*

74. Distance. Let P_1 be a point whose coordinates are (x_1, y_1), and let P_2 be a second point with coordinates (x_2, y_2). Since P_1 and P_2 are different points, we cannot have both the equalities $x_1 = x_2$ and $y_1 = y_2$ although we may have one or the other or neither. We shall discuss these three cases separately.

Case 1. Suppose $x_1 = x_2$; that is, P_1 and P_2 are on the same

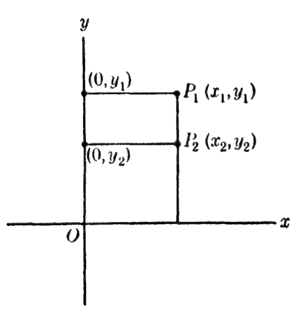


FIG. 73

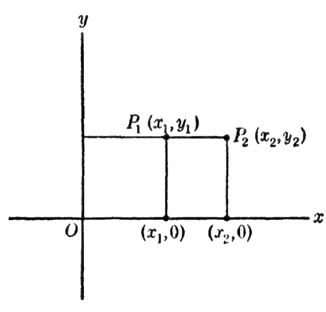


FIG. 74

vertical line. Then the distance P_1P_2 is equal to the distance between the points $(0,y_1)$ and $(0,y_2)$. (Why?) But by Theorem 1, section 71, this distance is $|y_1 - y_2|$ or $\sqrt{(y_1 - y_2)^2}$.

For example, the distance between the points $(1, -3)$ and $(1,2)$ is $|(-3) - 2| = |-5| = 5$.

Case 2. Suppose $y_1 = y_2$; that is P_1 and P_2 are on the same horizontal line. Then the distance P_1P_2 is equal to the distance

between the points $(x_1,0)$ and $(x_2,0)$. (Why?) But by section 71, this distance is $|x_1 - x_2|$ or $\sqrt{(x_1 - x_2)^2}$.

For example, the distance between the points $(-2,4)$ and $(1,4)$ is $|(-2) - 1| = |-3| = 3$.

Case 3. The only remaining possibility is that $x_1 \neq x_2$ and $y_1 \neq y_2$. Then the line P_1P_2 is oblique, that is, neither horizontal nor vertical. Draw a horizontal line through P_1 and a vertical line through P_2. These lines meet at a point M whose coordinates are (x_2,y_1); see Fig. 75 (the student should verify that our statements are true even when P_1 and P_2 are in positions other than those pictured).

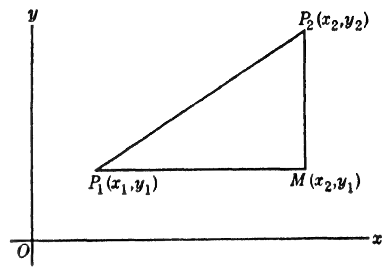


FIG. 75

Therefore by the two previous cases, P_1M is $\sqrt{(x_1 - x_2)^2}$ and MP_2 is $\sqrt{(y_1 - y_2)^2}$. Thus we have $(P_1M)^2 = (x_1 - x_2)^2$ and $(MP_2)^2 = (y_1 - y_2)^2$. Since angle M is a right angle we have $(P_1P_2)^2 = (P_1M)^2 + (MP_2)^2$ by the Pythagorean theorem. Therefore

$$(P_1P_2)^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2, \text{ or}$$

$$P_1P_2 = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \tag{1}$$

For example the distance between the points $(1,-1)$ and $(4,3)$ is

$$\sqrt{(1-4)^2 + (-1-3)^2} = \sqrt{(-3)^2 + (-4)^2} = 5.$$

Note that formula (1), which was derived from the hypothesis of Case 3, applies also to Cases 1 and 2. If (1) is applied to Case 1 where $x_1 = x_2$, the term $x_1 - x_2 = 0$ and (1) reduces to $\sqrt{(y_1 - y_2)^2}$; thus the formula (1) yields the correct result for Case 1.

Exercise. Verify that the formula (1) yields the correct result for Case 2.

We have proved the following theorem.

THEOREM 3. *If P_1 and P_2 are any two points with coordinates (x_1,y_1) and (x_2,y_2) respectively, the distance P_1P_2 between them is given by*

$$P_1P_2 = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

EXERCISES

Find the distance between each of the following pairs of points:

1. (3,7) and (– 2,– 5).
2. (0,0) and (– 3,4).
3. (– 2,2) and (2,2).
4. (0,2) and (– 1,3).
5. (2,3) and (2,– 7).
6. (3,– 1) and (– 7,– 1).
7. (0,0) and (– 4,0).
8. (0,0) and (0,– 6).

9. Show that the triangle whose vertices are (– 1,– 2), (3,– 2), (1,5) is isosceles. Is it also equilateral?

10. Find the lengths of the sides of the triangle whose vertices are (0,0), (0,– 5) and (– 12,0).

11. Let A, B, C be points whose coordinates are (– 2,3), (– 6,– 3), and (1,1) respectively. Use the converse of the Pythagorean theorem to show that ABC is a right triangle.

12. As in exercise 11, show that the points (– 4,– 3), (6,– 5), (– 3,2) are the vertices of a right triangle.

13. (*a*) Show that the points (3,1) and (– 3,1) are on a circle whose center is (0,5).

(*b*) What is the radius of this circle?
(*c*) Does the point (– 4,– 2) lie on this circle?
(*d*) Does the point (2,4)?
(*e*) The point (– 4,2)?

14. Find a point on the x-axis equidistant from the points (5,4) and (6,– 3). Hint: let the desired point be $(x,0)$ and find x.

15. Find a point on the y-axis equidistant from the points (3,1) and (4,– 6). Hint: let the desired point be $(0,y)$ and find y.

16. Show that the points (– 3,– 3), (3,5), and (6,9) lie in a straight line. Hint: the sum of the lengths of any two sides of a triangle must be greater than the length of the third side.

75. Midpoint of a line-segment. Let $P'(x',y')$ be the midpoint of the line-segment joining $P_1(x_1,y_1)$ and $P_2(x_2,y_2)$. We shall find expressions for the coordinates of P' in terms of the coordinates of P_1 and P_2.

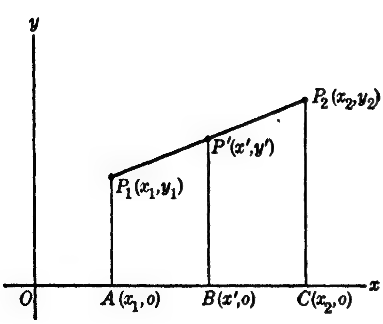


FIG. 76

Case 1. Suppose P_1P_2 is neither horizontal nor vertical. Draw vertical lines through the points P_1, P', P_2, intersecting the x-axis at $A(x_1,0)$, $B(x',0)$, $C(x_2,0)$ respectively. Clearly B is the midpoint of AC since if parallels intercept

equal segments on one transversal (P_1P_2) they intercept equal segments on any transversal (the x-axis). Thus, by Theorem 2, section 71, we have

$$x' = \frac{x_1 + x_2}{2}. \tag{1}$$

Similarly, by drawing three horizontal lines through P_1, P', P_2, we would obtain

$$y' = \frac{y_1 + y_2}{2}. \tag{2}$$

The proof of (2) is left to the student.

Case 2. Suppose P_1P_2 is horizontal. Then $y_1 = y' = y_2$. Hence $y' + y' = y_1 + y_2$ by substitution. Thus $2\,y' = y_1 + y_2$ and $y' = \frac{1}{2}(y_1 + y_2)$. Hence the same formula (2) applies in this case.

Formula (1) can be derived just as in Case 1.

Case 3. Suppose P_1P_2 is vertical. Then $x_1 = x' = x_2$. As in Case 2, we may prove that formula (1) holds in this case as well. The derivation of formula (2) in this case is the same as in Case 1.

Exercise. Complete the proofs in detail for each case.

Hence we have the following theorem.

THEOREM 4. *The coordinates (x',y') of the midpoint of the line-segment whose endpoints are (x_1,y_1) and (x_2,y_2) are given by*

$$x' = \frac{x_1 + x_2}{2}, \qquad y' = \frac{y_1 + y_2}{2}. \tag{3}$$

In other words, the abscissa of the midpoint is the average of the abscissas of the endpoints, and the ordinate of the midpoint is the average of the ordinates of the endpoints.

EXERCISES

1. Find the coordinates of the midpoint of the line-segment joining $(-1,-3)$ and $(7,11)$. Use the distance formula to verify that this midpoint is equidistant from the given points.

2. Find the midpoints of the sides of the triangle in:

(*a*) exercise 9, section 74; (*b*) exercise 10, section 74;
(*c*) exercise 11, section 74; (*d*) exercise 12, section 74.

3. Find the lengths of the 3 medians in each of the triangles of exercise 2.

4. For each of the triangles of exercise 2, show that the length of the line-segment joining the midpoints of any two sides of the triangle is equal to half the length of the remaining side.

5. The point $P(3,5)$ is on a circle whose center is $(-1,2)$. Find the coordinates of the other end of the diameter passing through P.

6. The points $(-2,6)$ and $(4,-2)$ are opposite ends of a diameter of a circle. Show that $(6,2)$ lies on the circle.

7. Write the coordinates of the midpoint of the line-segment joining (a,b) and (c,d).

76. Slope. We have already seen how the familiar geometric concepts of distance and midpoint of a line-segment may be treated algebraically. In this section we shall discuss a means of treating algebraically problems connected with the idea of the direction of a line, in particular, problems of parallel and perpendicular lines.

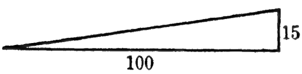


Fig. 77

If a hill rises 15 feet vertically for every 100 feet of horizontal progress, it is customary to say that it has a slope of 15/100. If it were steeper, rising, say, 30 ft. for every 100 feet of horizontal progress, it would have the larger slope 30/100. In Fig. 78, the line-segment P_1P_2 rises exactly $y_2 - y_1$ units (that is, the distance QP_2) while making horizontal progress of $x_2 - x_1$ units (that is, the distance P_1Q). Hence it is natural to make the following definition.

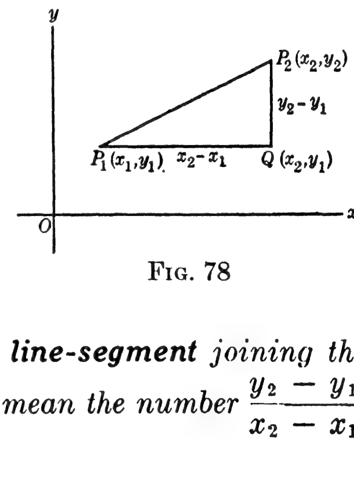


Fig. 78

Definition 1. *By the* ***slope of the line-segment*** *joining the points $P_1(x_1,y_1)$ and $P_2(x_2,y_2)$, we shall mean the number $\dfrac{y_2 - y_1}{x_2 - x_1}$, provided this number exists.*

Exercise. Prove that $\dfrac{y_2 - y_1}{x_2 - x_1} = \dfrac{y_1 - y_2}{x_1 - x_2}$. It follows that it does not matter which of the two given points we take as P_1 in applying the definition.

Example. The slope of the line-segment joining the points $(5,3)$ and $(8,1)$ is $\dfrac{1-3}{8-5} = \dfrac{-2}{3}$. If we had taken the second point as P_1, we would obtain $\dfrac{3-1}{5-8} = \dfrac{2}{-3} = -\dfrac{2}{3}$, just as before.

However, the slope of the line-segment joining (2,1) and (5,3) is $\dfrac{3-1}{5-2} = \dfrac{2}{3}$.

The geometric significance of our definition will be brought out by the next two theorems.

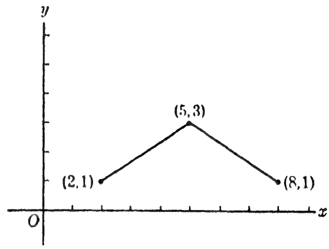


FIG. 79

THEOREM 5. *If a line-segment rises as we proceed from left to right, its slope is positive. If it sinks as we proceed from left to right, its slope is negative. If it is horizontal, its slope is zero. If it is vertical, it has no slope at all.*

Proof. Let $P_1(x_1,y_1)$ be to the left of $P_2(x_2,y_2)$. Then $x_1 < x_2$, or $x_2 - x_1$ is positive. If the segment P_1P_2 rises as we proceed from left to right, then $y_2 > y_1$, or $y_2 - y_1$ is positive (Fig. 78). Hence the slope is positive since it is obtained by dividing one positive number, $y_2 - y_1$, by another, $x_2 - x_1$.

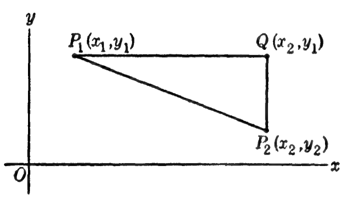


FIG. 80

If the segment sinks as we proceed from left to right then $y_2 < y_1$ and $y_2 - y_1$ is negative (Fig. 80). Hence the slope is negative since it is obtained by dividing a negative number, $y_2 - y_1$, by a positive number, $x_2 - x_1$.

By the exercise above, the slope is the same whether we take P_1 to be the left hand point or the right hand point.

If the segment is horizontal, then $y_2 = y_1$, or $y_2 - y_1 = 0$. Hence the slope is zero.

If the segment is vertical, then $x_2 = x_1$, or $x_2 - x_1 = 0$. Hence zero occurs in the denominator of the expression for the slope and the slope is therefore undefined for this case. This completes the proof.

The converses of the statements in Theorem 5 are also correct but we shall not prove them here.

THEOREM 6. *Let the line-segment P_1P_2 be oblique. Draw a horizontal line through P_1 and a vertical line through P_2, meeting at Q. Then the slope of the segment P_1P_2 is equal to* $+\dfrac{P_2Q}{P_1Q}$ or $-\dfrac{P_2Q}{P_1Q}$

according as the segment rises or sinks as we proceed from left to right.

Proof. Clearly the coordinates of Q are (x_2,y_1). Hence $y_2 - y_1$ is either the distance P_2Q or $-P_2Q$. Similarly $x_2 - x_1$ is either the distance, P_1Q or $-P_1Q$. Therefore, in any case, the slope is either $\dfrac{+P_2Q}{P_1Q}$ or $\dfrac{-P_2Q}{P_1Q}$. Which sign occurs has already been determined in Theorem 5.

THEOREM 7. *If two line-segments P_1P_2 and P_3P_4 are on the same non-vertical line, then their slopes are equal.*

Proof. Case 1. Suppose the line is horizontal. Then both slopes will be zero, by Theorem 5, and are consequently equal.

Case 2. Suppose the line is oblique. Draw horizontal and vertical lines as in Fig. 81. Angles Q and R are right angles. Angles P_1 and P_3 are equal since they are corresponding angles of parallel lines. Hence the right triangles are similar since two angles of one are equal to two angles of the other. Now, it is a theorem of plane geometry that if two triangles are similar, the ratio of the lengths of any pair of sides of one triangle equals the ratio of the lengths of the corresponding pair of sides of the other triangle. Hence $\dfrac{P_2Q}{P_1Q} = \dfrac{P_4R}{P_3R}$. By Theorem 6, these expressions are either the slopes of the segments P_1P_2 and P_3P_4 or the negatives of these slopes, according as the line rises or sinks as we proceed from left to right. In either case, the slopes are equal. This completes the proof.

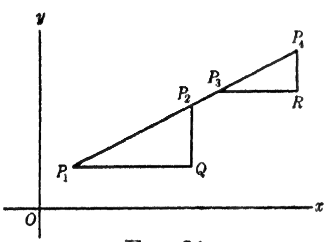


FIG. 81

By Theorem 7, the slope of any line-segment on a given line is the same as the slope of any other line-segment on the same line. This makes possible the following natural definition.

DEFINITION 2. *The **slope of a line** shall mean the slope of any line-segment on it.*

That the converse of Theorem 7 is false is apparent from the following theorem.

THEOREM 8. *If two non-vertical lines l and l' are parallel, then they have the same slope.*

Proof. Case 1. If the lines are horizontal, then their slopes are both zero.

Case 2. Suppose the lines are oblique. Since they are parallel, they either both rise or both sink as we proceed from left to right. Thus their slopes must have the same sign. Let P_1 and P_1' be the points where l and l' meet the x-axis, respectively. Let P_2 and P_2' be any other points on l and l', respectively. Let Q and Q' be the points where the vertical lines through P_2 and P_2' meet the x-axis, respectively. The angles P_1 and P_1' are equal, since l and l' are parallel, and Q and Q' are both right angles. Therefore, the right triangles are similar. Hence $\frac{P_2Q}{P_1Q} = \frac{P_2'Q'}{P_1'Q'}$. By Theorem 6, this implies that l and l' have the same slope.

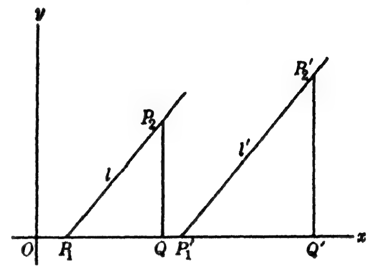


FIG. 82

THEOREM 9. *If two distinct* non-vertical lines l and l' have the same slope, then they are parallel.*

Proof. Case 1. If their slope is zero, the lines are both horizontal, and hence parallel.

Case 2. If their slope is not zero, the lines either both rise or both sink as we proceed from left to right, according as the slope is positive or negative. Draw right triangles P_1P_2Q and $P_1'P_2'Q'$ just as in the preceding theorem (Fig. 82). Since the slopes of l and l' are equal, we have $\frac{P_2Q}{P_1Q} = \frac{P_2'Q'}{P_1'Q'}$. Angles Q and Q' are equal, since they are both right angles. It is a theorem of plane geometry that, if one angle of a triangle equals one angle of another triangle, and the ratio of the sides including the first angle equals the ratio of the sides including the other, then the triangles are similar. Hence the triangles P_1P_2Q and $P_1'P_2'Q'$ are similar. Therefore, angle P_1 is equal to angle P_1'. Hence l and

* Two lines are called **distinct** if they are different, that is, if they do not coincide. Two distinct lines may very well have one point in common, however.

l' are parallel, because a pair of corresponding angles made by a transversal (the x-axis) are equal. This completes the proof.

If two line-segments have the same slope they are either on the same line or on parallel lines.

THEOREM 10. *Let l and l' be two oblique lines whose slopes are denoted by m and m' respectively. If $mm' = -1$, then the lines are perpendicular.*

Proof. Since $mm' = -1$, by hypothesis, m and m' are of opposite sign. Suppose, for example, m is positive. Then l rises and l' sinks as we proceed from left to right; hence they must intersect at some point A. Choose another point B on l. Draw a vertical line through B; this line must meet l' at some point C, since l' is not vertical. Draw a horizontal line through A meeting BC at D. By Theorem 6, $m = BD/AD$ while $m' = -DC/AD$. By hypothesis, $mm' = -1$ or $m = -1/m'$. Thus

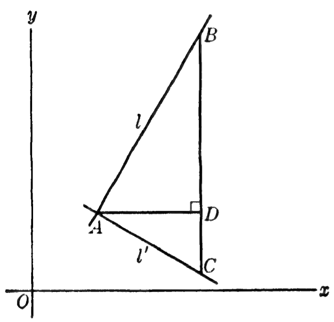


FIG. 83

$$\frac{BD}{AD} = \frac{-1}{-DC/AD}, \qquad \text{or} \qquad \frac{BD}{AD} = \frac{AD}{DC}.$$

Since angles BDA and ADC are right angles, it follows that the right triangles BAD and CAD are similar, by the theorem of geometry quoted in the preceding proof. Hence angle BAD is equal to angle C since corresponding angles of similar triangles are equal by definition. Now since triangle CAD is a right triangle, we have

$$\angle DAC + \angle C = 90°. \tag{2}$$

Substituting in (2) we have $\angle DAC + \angle BAD = 90°$. Therefore, $\angle BAC = 90°$. This is what we had to prove.

The converse of Theorem 10, is also true but we shall not prove it here.

Example. Let the points A, B, C, D have the coordinates (2,3), (4,7), (6,6), (4,2) respectively. Show that $ABCD$ is a rectangle. (The student should draw the figure.)

The slope of AB is 2; the slope of BC is $-\frac{1}{2}$; the slope of CD is

2; the slope of DA is $-\frac{1}{2}$. Since the slopes of AB and CD are equal, AB and CD are either parallel or on the same straight line; but they are not on the same straight line since the slope of BC is not also 2. Hence AB and CD are parallel. Similarly, AD and BC are parallel. Therefore, the figure is a parallelogram. But the slope of AB multiplied by the slope of BC yields -1. By Theorem 10, AB and BC are perpendicular. Therefore $ABCD$ is a rectangle.

EXERCISES

1. Show that the slope of the segment joining (1,2) and (3,8) is equal to the slope of the segment joining (4,10) and (8,22). Establish whether or not all four points are on the same line.

2. Show that the slope of the segment joining (1,2) and (2,6) is equal to the slope of the segment joining (4,14) and (9,34). Establish whether or not all four points are on the same line.

3. Find the slopes of the sides of the triangles in: (*a*) exercise 12, section 74; (*b*) exercise 11, section 74; (*c*) exercise 10, section 74; (*d*) exercise 9, section 74.

4. Find the slope of the lines joining the midpoints of the sides of each triangle in exercise 3. Establish in each triangle, that the line joining the midpoints of any two sides of the triangle is parallel to the third side.

5. Prove that the points $(-2,-2)$, $(-5,1)$, $(-4,5)$ and $(-3,-6)$ are vertices of a parallelogram.

6. Using Theorem 10, show that the triangles given in (*a*) exercise 11, section 74, (*b*) exercise 12, section 74, are right triangles.

7. Show that $A(1,6)$, $B(4,5)$, $C(1,-4)$ and $D(-2,-3)$ are the vertices of a rectangle.

8. Draw a line through the point (2,3) having (*a*) the slope 2; (*b*) the slope $\frac{1}{2}$; (*c*) the slope $-\frac{1}{2}$; (*d*) the slope -2.

9. Draw a line through the point (1,3) having (*a*) the slope 2/5; (*b*) the slope $-5/2$; (*c*) the slope $-2/5$; (*d*) the slope 5/2.

10. If A, B, C, D have the coordinates $(-2,0)$, $(2,4)$, $(6,0)$, and $(2,-4)$ respectively, show (*a*) that $ABCD$ is a square; (*b*) that the diagonals of this square are equal and perpendicular to each other; (*c*) that the diagonals bisect each other.

11. If A, B, C, D have the coordinates $(-1,0)$, $(3,4)$, $(7,-2)$, $(4,-3)$ respectively, show that the line-segment joining the midpoints of AB and BC is equal and parallel to the line-segment joining the midpoints of CD and DA.

12. State and prove the converse of Theorem 10.

77. Proofs of theorems. By analytic methods (that is, the use of coordinates) we can prove easily theorems in geometry which demand considerable ingenuity when attacked by the "synthetic" methods of the ancient Greeks. We shall take up some simple examples here. It will be useful to remember that we may choose our axes wherever we wish. We shall usually find it convenient to choose axes in such a way that as many coordinates as possible are zero.

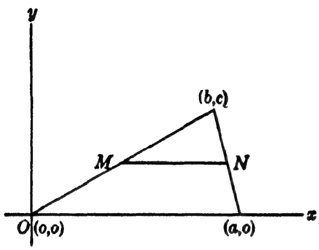


Fig. 84

Example. Prove that the length of the line-segment joining the midpoints of two sides of a triangle is half the length of the third side.

Proof. Choose the line of the third side as our x-axis and let one vertex be the origin. Then the coordinates of the other two vertices may be denoted by $(a,0)$ and (b,c) where $a > 0$. By Theorem 4, section 75, the midpoints M and N have coordinates $(b/2,c/2)$ and $\left(\frac{b+a}{2}, \frac{c}{2}\right)$ respectively. Hence, by Theorem 3, the distance $MN = \sqrt{\left(\frac{b+a}{2} - \frac{b}{2}\right)^2 + \left(\frac{c}{2} - \frac{c}{2}\right)^2} = \sqrt{\left(\frac{a}{2}\right)^2} = \frac{a}{2}.$ But the length of the third side is $a - 0 = a$. This proves the theorem.

Remark. If we had assigned definite numerical coordinates to the vertices we would have proven the theorem only for a particular triangle. By using unspecified literal coordinates we prove the theorem in all generality.

Note that our proof is much shorter than the proof you learned in high school, which involved making judiciously chosen construction lines, and a more or less ingenious argument. Here we have only to apply our formulas.

EXERCISES

1. Prove the theorem above, using the other pairs of sides in Fig. 84.

2. Prove the following theorems analytically, remembering to choose the axes in a convenient position:

(*a*) The diagonals of a rectangle are equal.

(*b*) The diagonals of a rectangle bisect each other.

(*c*) The midpoint of the hypotenuse of a right triangle is equidistant from the three vertices.

(*d*) Two medians of an isosceles triangle are equal.

(*e*) The diagonals of a square are perpendicular.

(*f*) The line-segments joining the midpoints of the sides of any quadrilateral in succession form a parallelogram.

(*g*) The line-segments joining the midpoints of opposite sides of any quadrilateral bisect each other.

(*h*) If the diagonals of a rectangle are perpendicular, then the figure is a square. (Hint: use the converse of Theorem 10, section 76.)

78. The graph of an equation. We have already seen that points in a plane may be identified with pairs of real numbers, and that certain geometric concepts, like distance, etc., may be studied by means of algebraic formulas. To see how other geometrical considerations may be reduced to algebraic problems, we need the following definition.

DEFINITION. *By the* ***graph*** *or* ***locus*** *or* ***curve of an equation*** *in two variables x and y we mean the set of (1)* ***all*** *those points, and, (2)* ***only*** *those points, whose coordinates satisfy* * *the equation.*

Example 1. Consider the equation $x + y = 5$. The point (2,3) is in the locus of this equation since $2 + 3 = 5$. That is, the pair of values $x = 2$, $y = 3$ satisfies the equation $x + y = 5$, or converts it into a true statement. So are the points (3,2), (1,4), $(-2,7)$, $(6,-1)$. But the point (1,3) is not in the graph since $1 + 3 \neq 5$.

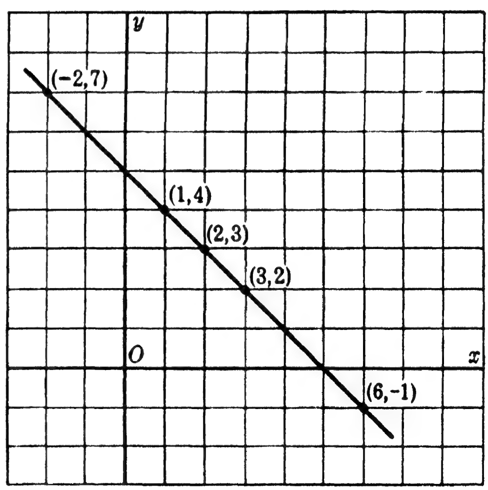


FIG. 85

Clearly we may substitute any value for x and obtain a corresponding value for y from the relation $y = 5 - x$. Many corresponding pairs of values may be tabulated as follows.

* A pair of values for x and y is said to **satisfy** an equation if the equation becomes a true statement when these values are substituted in it.

x	2	3	1	-2	6	$\cdots$
y	3	2	4	7	-1	$\cdots$

Marking these points on the graph (Fig. 85) we see that the graph seems to be a straight line.

Remark. Note that both conditions (1) and (2) are important. Thus the set of the 5 points which we actually plotted *alone* would satisfy condition (2) but not condition (1). On the other hand the set of all points on the line in Fig. 85 together with the point (1,3) would satisfy condition (1) but not condition (2).

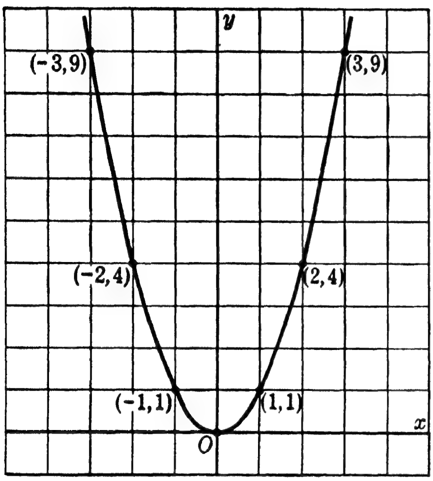


Fig. 86

To plot the graph of an equation, we usually substitute arbitrary values for x (or y) and calculate corresponding values of y (or x) from the equation. Many such substitutions must be made before we can form an idea of the shape of the curve. In joining the isolated points we have plotted by a "smooth" curve, we are assuming that the curve does not behave queerly between these points. The question of whether this is justified, and the question of how many points are necessary before the shape of the curve can be roughly determined, will be discussed in the next chapter. The coordinates of the points to be plotted are conveniently arranged in a table as in the following examples.

Example 2. Plot the graph of $y = x^2$. Substituting various numbers for x, we obtain the following table

x	0	1	-1	2	-2	3	-3	$\cdots$
y	0	1	1	4	4	9	9	$\cdots$

from which we get the graph in Fig. 86.

Example 3. Plot the graph of $x^2 + y^2 = 25$. We solve for y in terms of x as follows:

$$y^2 = 25 - x^2,$$
$$y = \pm\sqrt{25 - x^2}.$$

Substituting for x, we obtain the following table:

x	0	± 5	3	4	-3	-4	$+2$	-2	$\cdots$
y	± 5	0	± 4	± 3	± 4	± 3	$\pm\sqrt{21}$	$\pm\sqrt{21}$	$\cdots$

Clearly for any value of x greater than 5 or less than -5, the values of y will be imaginary. Hence no such points appear on the graph, since our coordinates are exclusively real numbers. We get the graph in Fig. 87.

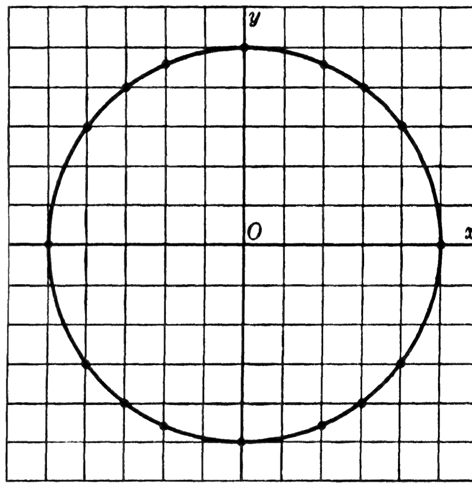


FIG. 87

Example 4. Plot the graph of $xy = 1$. Solving for y, we get $y = 1/x$. Clearly for $x = 0$, there is no point of the graph since $1/0$ is not defined. The following table may be obtained by substitution:

x	1	2	3	4	5	1/2	1/3	1/4	1/5	-1	-2	-3	$-1/2$	$-1/3$	$-1/4$	$\cdots$
y	1	1/2	1/3	1/4	1/5	2	3	4	5	-1	$-1/2$	$-1/3$	-2	-3	-4	$\cdots$

We get the graph in Fig. 88.

Example 5. Plot the graph of $9x^2 + 4y^2 = 36$. Solving for y we get $y^2 = \dfrac{36 - 9x^2}{4}$ or $y = \pm\dfrac{1}{2}\sqrt{36 - 9x^2}$. Substituting for x we get the following table:

x	0	± 1	± 2	$\cdots$
y	± 3	$\pm\dfrac{3\sqrt{3}}{2}$	0	$\cdots$

Clearly values of x greater than 2 or less than -2 yield imaginary values of y. Hence no such points appear on the graph (Fig. 89).

In general, the **locus** of all points satisfying certain given conditions means the set of all those points and only those points

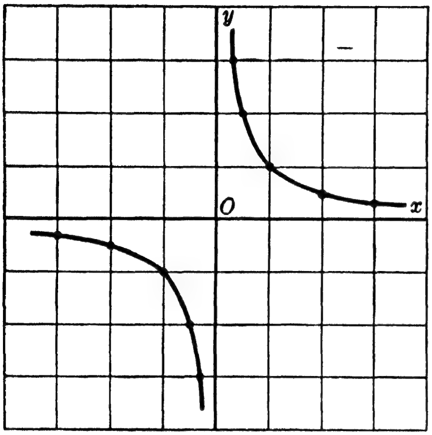


FIG. 88

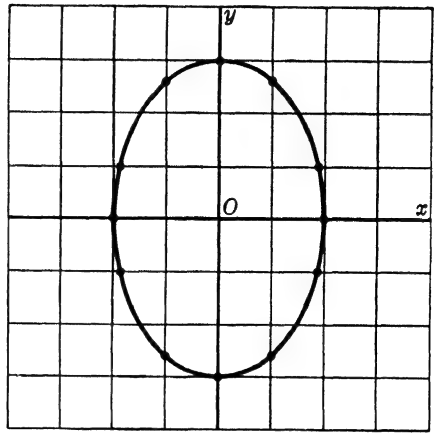


FIG. 89

which satisfy the conditions. For example, a *circle* is defined as the locus of all points in the plane whose distances from a fixed point, called the center, are equal to a given distance, called the radius. An **equation of a locus** means an equation satisfied by the coordinates of all the points of the locus and of no other points. Note that an equation may be altered by multiplying or dividing both sides of it by the same constant ($\neq 0$) without affecting the graph. For, if x and y are numbers for which the given equation is true, then the new equation must be true as well, since if equals are multiplied or divided by equals the results are equal. For example, if the point (x,y) satisfies the equation

(1) $$x + y = 5,$$

it also satisfies the equations

$$3x + 3y = 15,$$

obtained by multiplying both sides of (1) by 3, and

$$\frac{x}{2} + \frac{y}{2} = \frac{5}{2},$$

obtained by dividing both sides of (1) by 2, and conversely. Thus, while we may speak of *the* graph of an equation, we should

speak of *an* equation of a graph. However, it is usual to speak loosely of the equation of the graph, or locus. For example, an equation of the locus of all points 3 units above the x-axis is $y = 3$. Another equation of the same locus is $2y = 6$.

Let us now see how the geometric properties of lines, circles and other simple curves, may be studied by means of their equations.

EXERCISES

1. (*a*) Is the point (2,3) on the graph of the equation $3x - 2y = 0$? (*b*) The point (3,2)? (*c*) The point (4,6)?

2. (*a*) Does the curve of the equation $y = x^2 + 1$ pass through the point (3,10)? (*b*) The point (3, − 10)? (*c*) The point (− 3,10)?

Plot the graphs of the following equations:

3. $3x - 4y = 7$. **4.** $3x + 4y = 7$. **5.** $x^2 + y^2 = 100$.
6. $10y = x^3$. **7.** $y = 2x^3 + 3x^2 - 36x$. **8.** $xy = 4$.
9. $4x^2 + 9y^2 = 36$. **10.** $x^2 - y^2 = 9$. **11.** $y^2 = 10x$.
12. $10y = -x^2$. **13.** $y^2 = -10x$.

Write in simple form an equation of each of the following loci, and plot the graph:

14. The locus of all points 5 units to the right of the y-axis.

15. The locus of all points 3 units below the x-axis.

16. The locus of all points 2 units to the left of the y-axis.

17. The locus of all points whose abscissas and ordinates are of the same sign and which are equidistant from the x and y axes.

18. The locus of all points whose abscissas and ordinates are of opposite sign and which are equidistant from the x and y axes.

19. The locus of all points whose abscissas and ordinates are of the same sign and which are twice as far from the x-axis as from the y-axis.

20. Show that $x^2 - y^2 = 0$ is an equation of the locus of all points equidistant from the x and y axes. Plot the locus.

21. (*a*) Show that the values $x = 2$, $y = 3$ satisfy both the equations $2x + 3y = 13$ and $2x - y = 1$. Plot the graphs of both equations and discuss the geometric significance of the preceding statement.

(*b*) The values $x = 5$, $y = 1$ satisfy one of these equations but not the other. What is the geometric significance of this statement?

(*c*) The point (3,5) is on one of these graphs but not the other. What is the algebraic significance of this statement?

(*d*) The point (1,1) is on neither graph. What is the algebraic significance of this statement?

79. The equation of a straight line.

Case 1. Horizontal lines. Every point on a horizontal line has the same y-coordinate as every other point on it, and all points with this y-coordinate are on the line. Hence the equation of a horizontal line is of the form

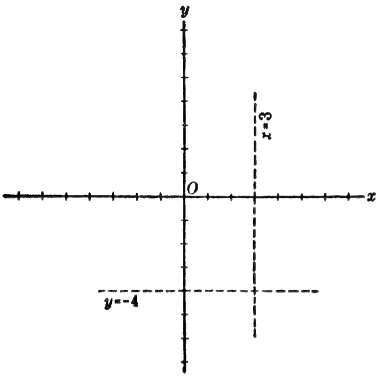


FIG. 90

(1) $$y = k$$

where k is a constant.

Case 2. Vertical lines. Similarly every vertical line has an equation of the form

(2) $$x = k$$

where k is a constant.

Case 3. Oblique lines. These lines have a slope different from zero. Consider the line having a given slope m and passing through a given point (x_1,y_1). Let (x,y) be any *other* point on the line. Then the line-segment joining (x,y) to (x_1,y_1) must have the slope m, or

(3) $$\frac{y - y_1}{x - x_1} = m.$$

Thus any point (x,y) on the line other than (x_1,y_1) satisfies the equation

(4) $$y - y_1 = m(x - x_1).$$

The point (x_1,y_1) itself clearly satisfies (4) as we see by direct substitution. Hence, every point on the line satisfies equation (4). Conversely, if (x,y) is any point, other than (x_1,y_1), satisfying (4), then (x,y) also satisfies (3); hence the segment joining (x,y) and (x_1,y_1) has the slope m and (x,y) is on the line. Thus (4) is an equation of the line passing through the point (x_1,y_1) with slope m. We have proved the following theorem.

THEOREM 11. *An equation of the line passing through (x_1,y_1) with slope m is given by (4). Horizontal and vertical lines have equations of the form (1) and (2) respectively.*

Note that equation (4) applies equally well to the case of horizontal lines, taking $m = 0$.

As a special case, consider the equation of the line passing

through a given point $(0,p)$ on the y-axis, with slope m. From (4) we see that its equation is $y - p = m(x - 0)$, or

$$y = mx + p. \tag{5}$$

? intercept on y axis

Hence we have the following theorem.

THEOREM 12. *An equation of the line passing through $(0,p)$ with slope m is given by (5).*

Suppose we want the equation of the oblique line passing through two given points (x_1,y_1) and (x_2,y_2). The slope of the line must be $\dfrac{y_2 - y_1}{x_2 - x_1}$ and the line passes through (x_1,y_1). Hence, by (4) the equation of the line is

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) \tag{6}$$

or

$$(y - y_1)(x_2 - x_1) = (y_2 - y_1)(x - x_1). \tag{7}$$

Note that the latter form of the equation may be applied to all lines through two given points, oblique or not; if the two points are on the same vertical line then $x_2 - x_1 = 0$ and (6) cannot be used while (7) can. Hence, we have the following theorem.

THEOREM 13. *An equation of the line through two points (x_1,y_1) and (x_2,y_2) is given by (7).*

In practice, it is unnecessary to use (7) for, given the coordinates of two points, one can get the slope and then use (4).

Examples. The equation of the line through $(2,-3)$ with slope 4 is $y + 3 = 4(x - 2)$, by (4). The equation of the line through $(0,2)$ with slope 1/3 is $y = \frac{1}{3}x + 2$, by (5). The line through the points $(2,3)$ and $(4,-5)$ is $(y - 3)(4 - 2) = (-5 - 3)(x - 2)$, or $4x + y = 11$, by (7). The line through $(3,1)$ and $(3,5)$ is $(y - 1)(3 - 3) = (5 - 1)(x - 3)$ or $x = 3$ by (7).

Since every line is either horizontal, vertical, or oblique, we have incidentally proven the following theorem.

THEOREM 14. *Every straight line has an equation of the first degree in x and y.*

EXERCISES

Write equations of the lines satisfying the following conditions:

1. Passing through (5,1) with slope 2.
2. Passing through $(-1,-2)$ with slope 1/2.
3. Passing through (3,2) with slope 0.
4. Passing through (5,2) parallel to the y-axis.
5. Passing through (2,3) and (5,6).
6. Passing through $(2,-3)$ and $(-1,2)$.
7. Passing through $(2,-3)$ and (2,5).
8. Passing through $(3,-2)$ and $(-1,-2)$.
9. Passing through (0,4) with slope 3.
10. Passing through (2,2) with slope 1.
11. Passing through $(-3,3)$ with slope -1.
12. Passing through (5,2) parallel to the x-axis.

13. If the line $y = 3x + p$ passes through (1,2), find p.
14. If the line $y = mx + 3$ passes through (1,2), find m.
15. If the line $y = mx + p$ passes through (1,2) and (3,5), find m and p.

80. The general equation of the first degree. We shall now prove the converse of Theorem 14.

THEOREM 15. *Every equation of the first degree in x and y has a straight line as its graph.*

Proof. Every equation of the first degree in x and y can be written in the form $ax + by = c$, a, b, and c being constants where not both a and b are zero. (Why do we make the latter restriction?) Either $b \neq 0$ or $b = 0$.

Case 1. Suppose $b \neq 0$. Then we may divide both sides of the equation $ax + by = c$ by b obtaining

$$\frac{a}{b}x + y = \frac{c}{b}$$

or

$$y = -\frac{a}{b}x + \frac{c}{b}.$$

But this is the equation of a line with slope $-\frac{a}{b}$ passing through the point $\left(0,\frac{c}{b}\right)$, by Theorem 12, section 79.

Case 2. Suppose $b = 0$. Then $a \neq 0$ since not both a and b can be zero. Then $ax + by = c$ becomes $ax = c$; since $a \neq 0$, we may divide through by a, obtaining

$$x = \frac{c}{a},$$

which by (2) of section 79, is the equation of a vertical line through $\left(\frac{c}{a},0\right)$.

This completes the proof.

An equation of the first degree is often called a **linear equation.** To plot such an equation we need compute the coordinates of only two points on the graph since we know that the graph is a straight line and two points determine a line. *The slope of the graph of a linear equation is most easily found by expressing the equation in the form $y = mx + p$; then the coefficient of x is the slope.*

Remark. Note that the line $y = -2x$ with slope -2 and passing through the origin has on it the points $(3,-6)$, $(2,-4)$, $(1,-2)$, $(0,0)$, $(-1,2)$, $(-2,4)$, $(-3,6)$, since $(-2)3 = -6$, $(-2)2 = -4$, $(-2)1 = -2$, $(-2)0 = 0$, $(-2)(-1) = 2$, $(-2)(-2) = 4$, $(-2)(-3) = 6$. This linear equation would not have a straight line for its graph were it not for the "rule of signs" for multiplication of signed numbers (section 21) which *defines* the products of signed numbers in the familiar way. This is perhaps the most important motivation for choosing the definition of multiplication of signed numbers as we did.

EXERCISES

Find the slope of the following lines, wherever possible, and plot the graph:

1. $2x - 3y = 6$. **2.** $2x + 5y = 10$. **3.** $3x = 1$. **4.** $2y = -10$.

5. $x + 2y = 0$. **6.** $3x - y = 0$.

7. What is the equation of the x-axis? the y-axis?

8. Write the equation of a line passing through $(-2,3)$ and parallel to the line $9x - 3y = 2$.

9. Write the equation of a line passing through $(3,2)$ and parallel to the line $6x + 3y = 5$.

10. Write the equation of a line passing through $(-2,3)$ and perpendicular to the line $9x - 3y = 2$.

11. A triangle has vertices $A(-2,0)$, $B(0,8)$, $C(4,2)$. Find the equations of (*a*) its sides; (*b*) its altitudes; (*c*) its medians; (*d*) its perpendicular bisectors.

12. Answer the questions of exercise 11 for the triangle with vertices $(-2,-4)$, $(2,12)$, $(6,-2)$.

Write in simple form an equation of each of the following loci, and plot the graph:

13. The locus of all points which are equidistant from the points $(1,2)$ and $(3,4)$. Is the point $(4,2)$ on this locus? The point $(6,-1)$?

14. The locus of all points equidistant from the points $(-1,2)$ and $(3,-4)$. Is the point $(7,3)$ on this locus? The point $(4,6)$?

15. The locus of all points whose distance from the origin is 5. Is the point $(3,4)$ on this locus? The point $(4,2)$?

16. The locus of all points whose distance from the point $(2,1)$ is 5. Is the point $(5,5)$ on this locus? The point $(6,4)$?

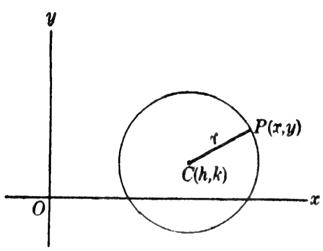


FIG. 91

81. The equation of a circle. By the **circle** with radius r and center at the point C we mean the set of all those points and only those points which are at a distance r from C. Let the coordinates of C be (h,k) and let the coordinates of any point P on the circle be (x,y). Then the distance PC must be r. Thus x and y must satisfy the equation

$$\sqrt{(x-h)^2+(y-k)^2} = r,$$

or

$$(x-h)^2+(y-k)^2 = r^2. \tag{1}$$

Conversely, if (x,y) satisfy (1), then the distance from (x,y) to (h,k) must be r ($-r$ cannot be regarded as a distance since it is negative). Hence (x,y) is on the circle. We have proved the following theorem.

THEOREM 16. *An equation of the circle with center at (h,k) and radius r is (1).*

Example. An equation of the circle with center at $(3,-2)$ and radius 5 is $(x-3)^2+(y+2)^2 = 25$.

EXERCISES

Write equations of the following circles:

1. With center at $(-1,3)$ and radius 4.

2. With center at $(2,-5)$ and radius 3.

3. With center at (0,0) and radius 2.

4. With center at (0,3) and radius 5.

5. With center at $(-2,0)$ and radius 2.

6. With center at (4,1) passing through (7,5).

7. With center at $(-2,-3)$ passing through (3,9).

8. With $(-2,3)$ and $(6,-7)$ as ends of a diameter.

Find the center and radius of the circle whose equation is:

9. $(x-2)^2+(y+5)^2=16$. **10.** $x^2+(y-3)^2=1$.

11. $(x-1)^2+y^2=9$. **12.** $x^2+y^2=25$.

Write in simple form an equation of each of the following loci, and plot the graph:

13. The locus of all points which can be the third vertex of a right triangle whose hypotenuse is the line-segment joining $(-5,0)$ and $(5,0)$. Is the point $(4,-3)$ on this locus? The point $(3,-4)$?

14. Given the points $A(0,10)$, $B(0,-10)$, and $P(x,y)$. Find the equation of the locus of all points P such that PA is perpendicular to PB. Is the point (4,5) on this locus? The point $(6,-8)$?

15. Given the points $A(-2,0)$, $B(2,0)$ and $P(x,y)$. Find the equation of the locus of all points P such that the sum of the squares of the distances PA and PB is equal to 58. Is the point $(4,-3)$ on this locus? The point (0,5)?

82. The general equation of degree two. In the last section we saw that the equation of every circle is quadratic; that is, of degree two. The converse proposition (namely, that the graph of every quadratic equation in x and y is a circle) is *false*. However, it *can* be proved that every quadratic equation in x and y has a graph which is a **conic section** (that is, a curve which can be obtained as the intersection of a plane with a cone) except for trivial exceptional cases.* Let us discuss these curves in further detail.

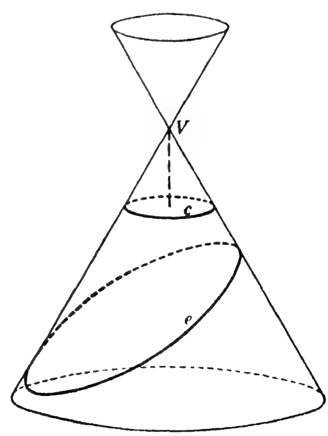


FIG. 92

Take a circle C (Fig. 92) in a horizontal plane and a point V, called the *vertex*, directly over the center of the circle. Consider all the lines, called *generators*, joining the vertex to the

* For example, the quadratic equation $x^2+y^2=-1$ has no locus at all since, x and y being real, neither x^2 nor y^2 can be negative and therefore no point can satisfy the equation.

points on the circumference C. The set of all the points on these lines is called a **cone** (more precisely, a *right circular cone*).

It is intuitively clear that every horizontal plane section of the cone is a circle, unless it be the vertex itself. A non-horizontal plane section cutting through two opposite generators is called an **ellipse** (e, Fig. 92). A plane parallel to a generator intersects the cone in a curve called a **parabola** (p, Fig. 93). A plane cutting both the upper and lower parts of the cone intersects the cone in a curve called a **hyperbola,** a curve which falls into two branches (h, Fig. 93).

Examples 2, 3, 4, 5 of section 78 are a parabola, circle, hyperbola, and ellipse respectively. These are the only types of conic sections, save for trivial special cases.*

The conic sections were studied by the ancient Greeks by "synthetic" methods like those you used in high school. A splendid book on the subject was written by Apollonius of Perga (about the 3rd century B.C.) in which a wealth of knowledge about the conic sections was obtained with great ingenuity. The analytic methods of Descartes enable us to study these curves with ease by means of quadratic equations, and furthermore to discover algebraically many theorems about them that the Greeks had never suspected. It is interesting to note that the Greeks studied the geometry of the conic sections merely as a beautiful chapter of pure mathematics, and had no thought of applications. About 1800 years later these curves were found to have the most important practical applications in physical science. It is, in fact, difficult to exaggerate their importance. Thus the conic sections provide striking evidence for the thesis that one can never tell when pure science may find unsuspected

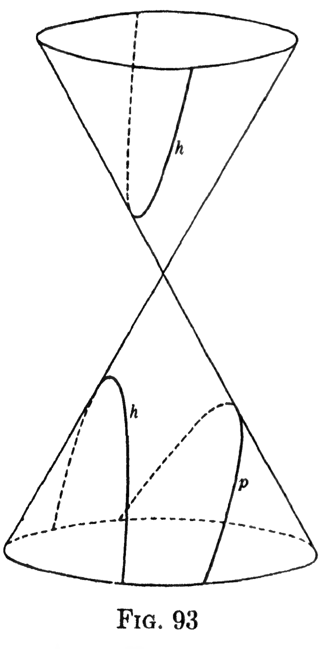


FIG. 93

* If the cutting plane passes through the vertex of the cone, the intersection may consist of a pair of straight lines, or a single straight line, or a single point.

practical applications. We shall sketch briefly a few of the many connections in which the conic sections are important.

The path of a projectile, like a baseball or a cannon ball, thrown obliquely, is a parabola, neglecting the influence of wind, etc.

Searchlight reflectors are parabolic surfaces obtained by rotating a parabola about its axis (the x-axis in Fig. 94). Associated with a parabola is a point called its *focus*. This point (F in Fig. 94) has the property that all light rays emanating from it will be reflected in the same direction, thus concentrating the beam of the searchlight in that direction and preventing it from spreading. For the same reason the mirrors in reflecting telescopes used in astronomy are parabolic. For the light rays entering the telescope from a distant star are, practically, parallel. Hence the parabolic reflectors concentrate all these rays at the focus.

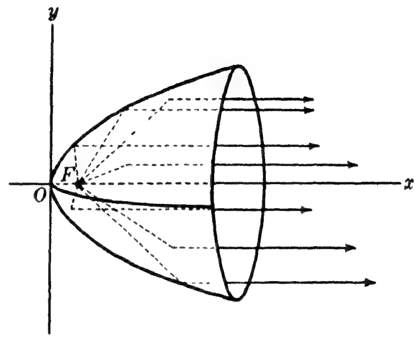


FIG. 94

The cable of a suspension bridge is, under certain circumstances, a parabola. Arches are sometimes parabolic in shape.

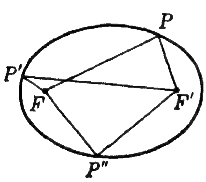


FIG. 95

Associated with an ellipse there are two points within it called the *foci* of the ellipse (F and F' in Fig. 95). The ellipse has the property that, for any point P on it whatever, the sum of the distances $PF + PF'$ is always the same. For example, in Fig. 95, $PF + PF' = P'F + P'F' = P''F + P''F'$. This property enables one to construct an ellipse easily by looping a string about two thumbtacks, inserting a pencil point so that the string is held taut, and moving the point about. Thus one gets an ellipse with the two tacks as foci.

The orbits of the planets are ellipses with the sun at one focus. This alone is sufficient to explain the overwhelming importance of ellipses since the 17th century when Kepler and Newton did their monumental work in astronomy.

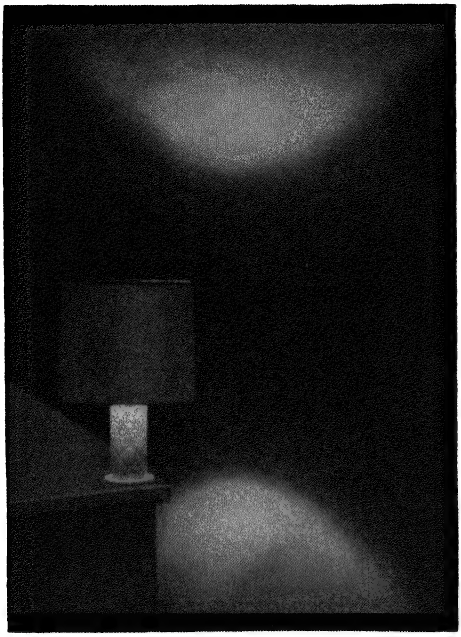


Fig. 96

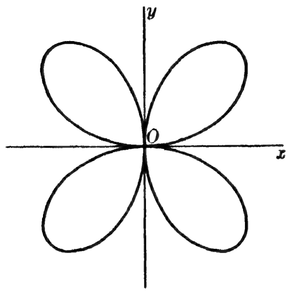
The ellipse has interesting reflecting properties. It can be shown that all rays of light, or sound waves, emanating from one focus must collect again at the other focus. This explains the amusing phenomenon of the whispering gallery. If the walls or ceiling of a hall are elliptical, then a whisper at one focus may not be audible at all at a nearby place but may nevertheless be audible far off at the other focus, since all the individually weak sound waves scattering over the room will gather together there (Fig. 95). One such whispering gallery exists in Statuary Hall, Washington, D. C.

The orbits of some meteors are parabolas while others are branches of hyperbolas. Many scientific laws are expressed as quadratic equations involving two variable quantities, and hence their graphs are conic sections. But the applications of the conic sections in modern science are too numerous to list.

Fairly good approximations of the various conic sections may be seen by shining a flashlight on a wall in an otherwise dark room and varying the angle at which the flashlight is held. This is so because the light emanates from the circular opening of the flashlight in an approximate cone and the wall plays the role of the intersecting plane. Both branches of a hyperbola may be seen when a lamp with a cylindrical shade is placed near a wall (Fig. 96).

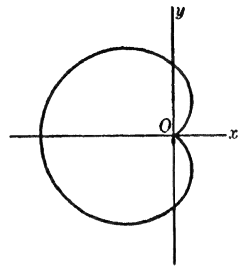
After discovering the connection between conic sections and equations of the second degree in x and y, the mathematician's desire to obtain general results naturally leads him to investigate the geometric properties of the graphs of equations of higher degree. This study is called the *theory of algebraic curves* and is a

very difficult and intricate branch of modern mathematics in which research is still being done. Many parts of modern mathe-



Four-leaved clover. $(x^2+y^2)^3=4x^2y^2$

FIG. 97



Cardioid: $(x^2+y^2+x)^2=x^2+y^2$

FIG. 98

matics, like this one, arise from elementary problems which come up naturally in elementary work.

EXERCISES

Plot the following curves:

1. The parabola $y = 10 - \frac{1}{2}x^2$.
2. The parabola $y = \frac{1}{2}x^2 - 10$.
3. The hyperbola $xy = 10$.
4. The hyperbola $x^2 - y^2 = 1$.
5. The ellipse $4\,x^2 + 25\,y^2 = 100$.
6. The ellipse $25\,x^2 + 4\,y^2 = 100$.

Write in simple form an equation of each of the following loci and plot the graph:

7. The locus of all points P whose distances from the line $x = -2$ and from the point (2,0) are equal.

8. The locus of all points P such that the distance from P to the point $(-4,0)$ is equal to $\frac{4}{5}$ of the distance from P to the line $x = -\frac{25}{4}$.

9. The locus of all points P such that the distance from P to the point (5,0) is equal to $\frac{5}{4}$ of the distance from P to the line $x = \frac{16}{5}$.

83. Common chord of two intersecting circles. Consider any two circles

$$(1)\qquad (x-h)^2+(y-k)^2=r^2$$

$$(2)\qquad (x-H)^2+(y-K)^2=R^2,$$

intersecting at two points. Form a new equation by subtracting (2) from (1), obtaining

(3) $(x - h)^2 - (x - H)^2 + (y - k)^2 - (y - K)^2 = r^2 - R^2,$

which can be simplified to the form

(4) $(2H - 2h)x + (2K - 2k)y = r^2 - R^2 - h^2 - k^2 + H^2 + K^2.$

This is a linear equation in x and y (the other letters representing constants). Suppose (x_1,y_1) is a point of intersection of the two circles; then (x_1,y_1) must satisfy both (1) and (2); hence it must satisfy (3) since if equals are subtracted from equals the results are equal. Therefore it must satisfy (4). By the same argument, the second point of intersection (x_2,y_2) must satisfy (4). But (4) is a linear equation and its graph is therefore a straight line. Now, two points determine a straight line and we already know that our two points of intersection are on this line (4). Hence (4) is an equation of the line joining the two points of intersection of the given circles. This line is called the **common chord** of the two circles. We have proved the following theorem.

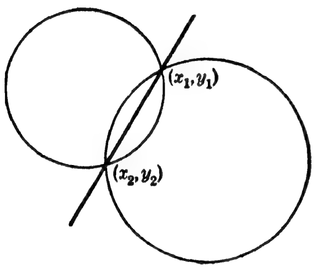


FIG. 99

THEOREM 17. *An equation of the common chord of two intersecting circles may be obtained by subtracting the equation of one circle from the equation of the other circle (provided they are written so that the coefficients of x^2 and y^2 are the same in both equations) and simplifying.*

EXERCISES

Find an equation of the common chord of the given circles and plot:

1. $(x - 3)^2 + (y - 5)^2 = 25$
$(x - 5)^2 + (y - 3)^2 = 25.$

2. $(x + 1)^2 + (y - 2)^2 = 16$
$(x - 1)^2 + (y + 2)^2 = 9.$

3. $x^2 + y^2 = 16$
$(x + 2)^2 + (y - 2)^2 = 9.$

4. $(x + 2)^2 + (y - 1)^2 = 5$
$x^2 + y^2 = 10.$

84. Intersections of two curves. Consider the graphs of two equations. By a point of intersection of the two curves we mean

a point on both of them. If (x_1,y_1) is such a point, then $x = x_1$ and $y = y_1$ must satisfy both equations because the point is on both curves. Conversely, if $x = x_1$ and $y = y_1$ are real numbers which satisfy both equations, then (x_1,y_1) is a point on both curves. Thus *the coordinates of the points of intersection of two curves are merely the real common solutions of the two equations; to find the points of intersection we have only to solve the two equations simultaneously.*

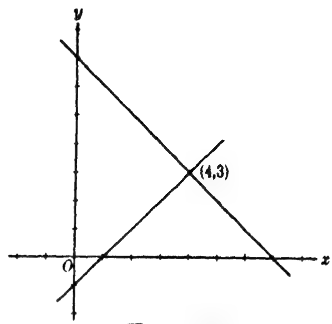


FIG. 100

Example 1. Let the curves be straight lines: say $x + y = 7$ and $x - y = 1$. From the first equation we obtain $y = 7 - x$. Substituting this in the second equation, and solving as in section 39, we have $x = 4$ and $y = 3$. Hence (4,3) is the point of intersection. The student should work out the details.

Example 2. Consider a straight line and a conic section: say

(1) $$x - y = 2, \text{ and}$$

(2) $$x^2 + y^2 = 34.$$

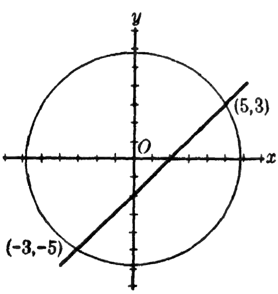


FIG. 101

From (1) we obtain $y = x - 2$. Substituting this in (2), and solving as in section 40, we find $x = 5$ or $x = -3$. For $x = 5$, we obtain $y = 3$ from (1); for $x = -3$ we obtain $y = -5$ from (1). Hence (5,3) and $(-3,-5)$ are the points of intersection. The student should complete the solution in detail.

Example 3. To find the intersections of two circles we have only to find the equation of their common chord (Theorem 17, section 83) and find the points of intersection of this line with one of the circles as in example 2. Thus, the circles

(3) $$(x - 2)^2 + y^2 = 5$$

(4) $$x^2 + (y - 1)^2 = 10$$

have the common chord

(5) $$2x - y = 4.$$

The points of intersection of this line with the circle (3) are found

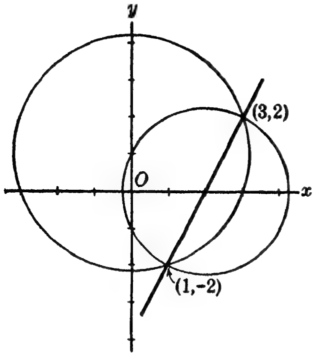


Fig. 102

to be (3,2) and (1, − 2). Hence these are the points of intersection of the two given circles. The details of the solution should be worked out by the student.

The general problem of finding the points of intersection of any two curves with equations of arbitrary degree is a very natural one to think of but difficult to answer. See section 40.

EXERCISES

Find all the points of intersection of the following curves and plot:

1. $2x - 3y = 4$
$3x + 4y + 11 = 0.$

2. $3x - 2y = 5$
$x + 5y = 1.$

3. $2x - 5y = 3$
$3x + y = 4.$

4. $y^2 = 4x$
$x + y = 3.$

5. $y^2 = 4x$
$y = 2.$

6. $y^2 = 4x$
$x = 1.$

7. $x^2 + y^2 = 25$
$4x - 3y = 0.$

8. $xy = 12$
$x + y = 7.$

9. $x^2 + y^2 = 41$
$x + y = 9.$

10. $x^2 - y^2 = 5$
$3x + y = 11.$

11. $y^2 = x$
$y - 4x + 3 = 0.$

12. $(x - 3)^2 + (y - 4)^2 = 4$
$x + y = 16.$

13. $(x - 2)^2 + (y - 5) = 0$
$y = x + 1.$

14. $(y + 1)^2 = x - 2$
$y + 12 - 4x = 0.$

15. $(x + 2)^2 + y^2 = 5$
$x^2 + (y - 1)^2 = 10.$

16. $(x + 1)^2 + (y + 1)^2 = 5$
$(x + 3)^2 + y^2 = 10.$

17. $(x + 2)^2 + y^2 = 5$
$x^2 + (y + 1)^2 = 10.$

18. $(x + 2)^2 + (y - 1)^2 = 5$
$x^2 + y^2 = 10.$

19. Plot the graphs of $x + y = 3$ and $x + y = 7$. Explain the geometric interpretation of the fact that these equations are incompatible (see section 39).

20. Plot the graphs of $x + 2y = 3$ and $2x + 4y = 6$. Explain the geometric interpretation of the fact that these equations are dependent or equivalent (see section 39).

21. In the triangle whose vertices are (0,0), (6,0) and (2,4):

(*a*) Find the coordinates of the midpoints of the sides;
(*b*) Find the equations of the medians;
(*c*) Find the point of intersection of two of the medians;
(*d*) Show that the three medians are concurrent;
(*e*) Show that the distance from each vertex to the point of intersection of the medians is 2/3 of the length of the corresponding median.

22. In the triangle of exercise 21,

(*a*) Find the equations of the three altitudes;
(*b*) Show that they are concurrent.

23. In the triangle of exercise 21,

(*a*) Find the equations of the three perpendicular bisectors of the sides;

(*b*) Show that they are concurrent.

(*c*) Show that the point of intersection of the altitudes (found in exercise 22), the point of intersection of the medians (found in exercise 21), and the point of intersection of the perpendicular bisectors all lie on the same straight line.

24. If A, B, C have the coordinates (2,1), (5,4), (7, − 3), respectively, find the coordinates of:

(*a*) the vertex D of the parallelogram $ABCD$.

(*b*) the vertex E of the parallelogram $ABEC$.

(*c*) the vertex F of the parallelogram $AFBC$.

85. Translation of axes. Invariants. All our work has been based on a definite set of axes (that is, a definite coordinate system) although we have at times allowed ourselves to choose these axes in a convenient position. Suppose, now, that we change from our given coordinate system to a new coordinate system whose origin is at the point (2,3) and whose axes are parallel to the old axes, and have similar positive directions. Call the old system the xy-system (Fig. 103), and the new one the $x'y'$-system. A change of coordinate system where the new axes are parallel to the old, with similar positive directions, is called a **translation** of axes. Consider the point P (Fig. 103) whose coordinates are (3,9) in the xy-system; clearly its coordinates in the $x'y'$-system are (1,6). Similarly the point Q whose coordinates in the xy-system are (6,5) has the coordinates (4,2) in the $x'y'$-system. Thus the values of the coordinates of a point depend upon where the pair of axes or "frame of reference" is located. However, consider the distance PQ. Using the xy-system,

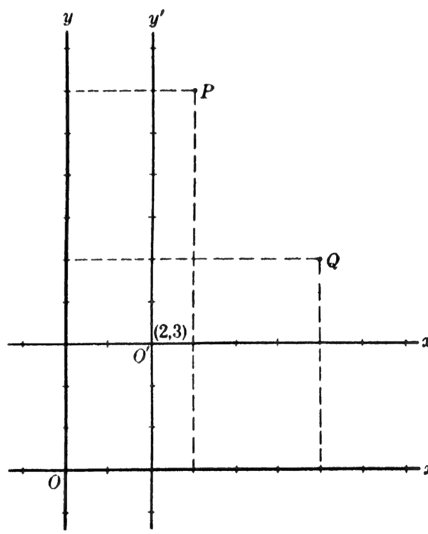


FIG. 103

$$PQ = \sqrt{(3-6)^2 + (9-5)^2} = \sqrt{(-3)^2 + (4)^2} = 5;$$

and using the $x'y'$-system,

$$PQ = \sqrt{(1-4)^2 + (6-2)^2} = \sqrt{(-3)^2 + (4)^2} = 5.$$

This suggests that, while the coordinates of a point change when we make a translation of axes, the distance between the two points remains unchanged or **invariant**; that is, the expression for the distance gives the same value in either system.

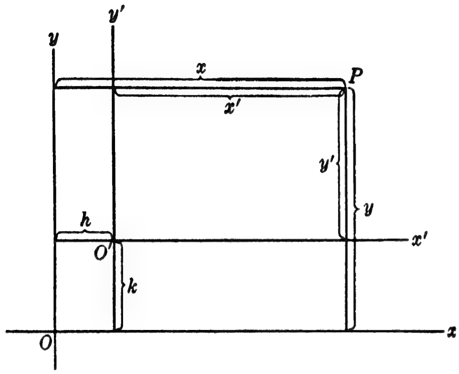


Fig. 104

In general, if we translate our axes so that the new origin is at a point whose old coordinates are (h,k), then for any point P with coordinates (x,y) in the old system and coordinates (x',y') in the new system, we have (Fig. 104)

$$(1) \qquad \begin{cases} x = x' + h \\ y = y' + k \end{cases}, \text{ or } \begin{cases} x' = x - h \\ y' = y - k. \end{cases}$$

Thus the coordinates of an individual point are altered by the translation. Let P_1 and P_2 be points whose old coordinates are (x_1,y_1) and (x_2,y_2), and whose new coordinates are (x_1',y_1') and (x_2',y_2'), respectively. Then by (1) we have

$$(2) \qquad \begin{array}{ll} x_1 = x_1' + h, & x_2 = x_2' + h, \\ y_1 = y_1' + k, & y_2 = y_2' + k. \end{array}$$

In the xy-system the distance

$$(3) \qquad P_1P_2 = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2},$$

and in the $x'y'$-system the distance

$$(4) \qquad P_1P_2 = \sqrt{(x_1' - x_2')^2 + (y_1' - y_2')^2}.$$

Substituting the values given by (2) in (3) we obtain

$$\begin{aligned} \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} & \\ &= \sqrt{(x_1' + h - x_2' - h)^2 + (y_1' + k - y_2' - k)^2} \\ &= \sqrt{(x_1' - x_2')^2 + (y_1' - y_2')^2}. \end{aligned}$$

This proves that the value of the expression called the distance formula is unaltered by a translation of the axes; that is, it yields the same numerical value no matter in which of the two coordi-

nate systems we apply it. With a little more technical knowledge than we assume here, it can be proved that the distance formula also remains invariant or unaltered if we rotate our axes (Fig. 105). On the other hand, the numerical value of the distance formula would not remain unchanged if we chose a different unit of length. Thus the points (2,0) and (4,0) have a distance of 2; but if we doubled the size of the unit of length they would have coordinates (1,0) and (2,0), respectively, and their distance would be 1. But even a change of unit will leave the ratio of two lengths invariant.

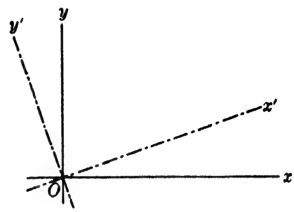


FIG. 105

From one point of view, the study of geometry is the study of

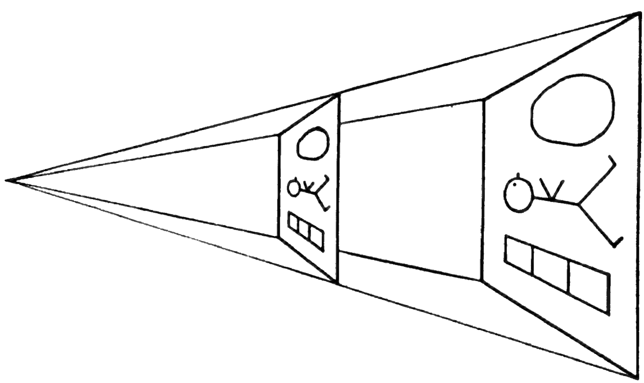


FIG. 106

those properties of figures which remain invariant under various kinds of changes. Thus distances (as we have proved), areas, angles, etc., are properties which remain invariant when translations (or rotations) of axes are made. The property of two triangles being similar remains invariant even under changes of units of length, as does the size of an angle, while length and area do not. Elementary geometry, considered from this point of view, is the study of those properties of figures which remain invariant under the types of changes or "transformations" mentioned above.

A figure on a flat piece of moving picture film is said to be *projected* into its image on a flat screen which may or may not be parallel to the film (Fig. 106). If one figure can be obtained from

another by a number of such projections, the second is said to have been transformed into the first by projection. The study of those properties of figures which remain invariant or unchanged by projections is called **projective geometry.** Clearly the property of being a circle is not invariant under projections, as may be seen from section 81. But the property of having a

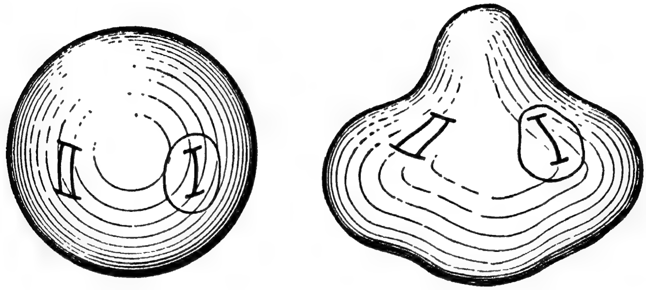


Fig. 107

second degree equation (that is, of being a conic section) is invariant under projections. Thus a circle can be projected into an ellipse or a parabola, etc., but not into a curve whose equation is of the third degree, say. A classic work on projective geometry was written by the French mathematician Poncelet (1788–1867) while he was a prisoner of war in Russia.

A very general kind of geometry, developed largely in the last 50 years is called **topology** or **analysis situs.** This subject may be partially described as the study of those properties of figures which remain invariant when the figure is deformed in any "continuous" way. Thus, if we imagine our figures to be made of rubber, we may think of them as being folded, stretched, bent, crumpled, and deformed in any way at all as long as they are not punctured or torn. It may be difficult to imagine that such drastic changes will leave any properties of the figure unchanged or invariant. But there are some. Thus a simple property of a spherical surface like a rubber balloon is that if we cut it along any "closed" curve like a circle, the surface will fall apart into two pieces (Fig. 107). This property will be preserved under deformations of the sphere. Notice that this property of falling into two pieces when cut along a closed curve is not possessed by a doughnut or inner tube; thus if a doughnut is cut along the circle C indicated in Fig. 108, it still hangs together in one piece. The

property of being inseparably linked as the two links of a chain in Fig. 109, is unchanged no matter how we deform continuously the shapes of the links. Topology has been called the science of carelessly drawn figures, since a topologist would not distinguish between the two figures in Fig. 109, for example, because he is interested only in properties which are preserved when one of these figures is deformed into the other. The four-color problem, the problem of the seven bridges of Königsberg, and the problem of 3 houses and 3 wells discussed in Chapter VIII are problems of topology since the actual rigid shape of the figures in question may be deformed continuously without affecting the problem. The study of topology is one of the most recent kinds of geometry to be developed, and has proved to be of importance even in application to physical science.

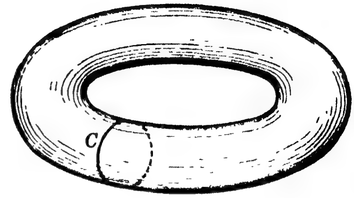


FIG. 108

The idea of invariants occurs in physical science itself. Our measurements of the position and velocity of a star, say, depend on the axes or frame of reference to which we refer its coordinates. This depends on the position and motion of the observer, among other things. But we would like our "laws of nature" to be independent of the observer. Hence we would like "laws of nature" to be expressible as formulas which are invariant under changes of frame of reference. Of course, in a broad sense, the task of the scientist is the search for invariant relationships under varying conditions; it is the search for permanence in a changing world.

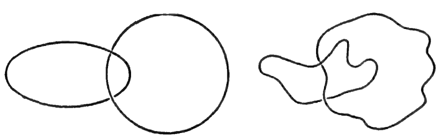


FIG. 109

86. Three-dimensional or solid geometry. Just as we associate a pair of coordinates with a point in a plane, we can associate a triplet of coordinates with a point in space. Taking three mutually perpendicular lines as x-axis, y-axis, and z-axis, respectively, we locate the point whose coordinates are $(2,3,-4)$ by proceeding 2 units in the direction of the positive x-axis, 3 units in the direction of the positive y-axis, and 4 units in the direction

of the negative z-axis (Fig. 110). Because each point in space is thus identified by means of three coordinates, the geometry of space is called 3-dimensional. It can be proved that the distance between two points (x_1,y_1,z_1) and (x_2,y_2,z_2) is

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2};$$

that the set of all points satisfying an equation of the first degree, such as $x + 2y - 3z = 4$, constitutes a plane; that the set of all points satisfying the equation $x^2 + y^2 + z^2 = 25$ constitutes the surface of a sphere of radius 5 with center at the origin; and so on. Thus the study of solid geometry, which requires considerable ingenuity when attacked by methods of the ancient Greeks, is reduced to the study of algebra with three variables, a subject which can be studied systematically with much less difficulty.

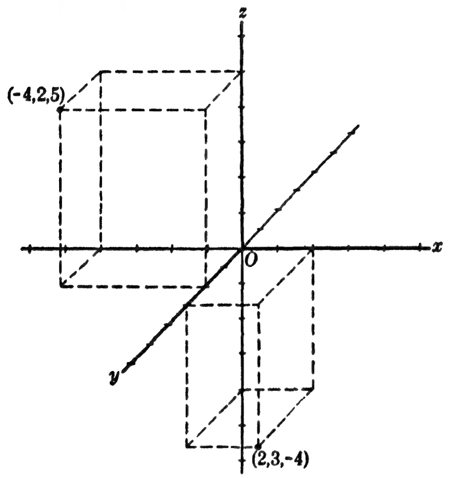


FIG. 110

87. Analytic geometry as a concrete interpretation of the postulates for geometry. We have said that any theorem that can be proved in geometry by means of the methods of the ancient Greeks can also be proved by means of coordinates. How can we be sure that this is so? This question can be answered simply by approaching analytic geometry from a slightly different point of view.

Let us recall the postulates for geometry given in Chapter II. We considered the words "point" and "line" as undefined terms and we assumed about these undefined terms such properties as the following:

POSTULATE 1. *Given two distinct points, there is at least one line containing them.*

The postulates and their logical consequences, the theorems, constitute the abstract mathematical science known as geometry.

As we pointed out in Chapter II, we usually give a concrete interpretation of this abstract mathematical science by assigning the meanings "dot" and "streak" to the undefined terms "point" and " line"; thus we interpret the whole geometry as applying to the diagrams we draw. But let us now give a different concrete interpretation to our abstract mathematical science, geometry, by assigning different meanings to the undefined terms "point" and "line," as follows. Let "point" mean a pair of real numbers (x,y). Let "line" mean a set of all those and only those "points" (that is, pairs of real numbers) which satisfy a given equation $ax + by = c$ of the first degree, where a, b, and c are all real constants and not both a and b are zero. It is not difficult to verify that the postulates become true statements when these meanings are substituted for the undefined terms.

For example, let us verify postulate 1. Let (x_1,y_1) and (x_2,y_2) be any two distinct given "points." We have only to show that there exists at least one equation of the first degree in x and y which is satisfied by both of these given pairs of numbers. But

$$(y - y_1)(x_2 - x_1) = (x - x_1)(y_2 - y_1)$$

is such an equation (see Theorem 13, section 79), for it is clearly satisfied by both $x = x_1$, $y = y_1$ and $x = x_2$, $y = y_2$, as we see by direct substitution. That the coefficients of x and y in this equation cannot both be zero follows at once from the hypothesis that (x_1,y_1) and (x_2,y_2) are distinct points, so that we can have one or the other or neither of the equalities $x_1 = x_2$ and $y_1 = y_2$ but not both.

Similarly all the other postulates can be verified algebraically if we interpret the undefined terms as having the above meanings. But if all the postulates are true statements for these meanings then so are all the theorems, for the theorems follow from the postulates by pure logic. Hence, analytic geometry may be regarded as a concrete interpretation of the abstract mathematical science discussed in Chapter II.

88. Geometry of more than three dimensions. In the last section we indicated how we might set up 2-dimensional (plane) analytic geometry by defining "point" to mean "pair of real numbers." Similarly we might set up 3-dimensional (solid) geometry by defining "point" to mean "triplet of real numbers."

But there is now nothing to stop us from defining "point" to mean a "set of 4 real numbers," like (x,y,z,u), or a "set of 17 real numbers," or a "set of n real numbers" where n is any natural number. If we did this we would be beginning the study of geometry of 4 or 17 or n dimensions. Thus, developing the subject in ways analogous to our study of 2-dimensional analytic geomerty, the study of 4- or 17- or n-dimensional geometry is nothing more than the study of algebra with 4 or 17 or n variables. This presents little difficulty. That is all there is, essentially, to the study of geometry of many dimensions.* The use of geometric terminology, such as calling a set of 17 real numbers a "point," is merely a convenience. As you see, there is nothing mysterious about this unless, indeed, you try to imagine, in some mysterious way, what 4-dimensional or 17-dimensional space looks like pictorially. Even mathematicians, whose eyes are constructed in the same way as yours, are unable to do that; in fact, there is no reason why they should want to, and they don't. Our visual spatial intuition is 3-dimensional and, indeed, even in 3-dimensions, our intuition cannot be trusted without the aid of reasoned ideas of perspective, etc.

If you now understand that "point in space of 5 dimensions" is taken to mean a set of 5 real numbers and "space of 5 dimensions" is taken to mean merely the totality of all such "points" or sets of 5 real numbers (x,y,z,u,v), you may fairly ask, "How can such a 5-dimensional geometry be applied practically?" The answer is that it may be very convenient to represent as a "point" in 5-dimensional space anything that can be specified by a set of 5 real numbers. For example, if we were making a statistical study of people with respect to their height, weight, age, blood pressure, and income, and were interested in no other properties, we might specify each person by a set of 5 real numbers (h,w,a,b,i); it might then be convenient to represent each person as a "point" in 5-dimensional space.

This kind of thing has actually been found to be of the greatest convenience in the study of dynamics. For example, consider 3 billiard balls moving on a table. At any instant, the position of each ball can be specified by its two coordinates. Thus the "position" of the dynamical system consisting of the 3 moving balls

* The subject may be approached from many different viewpoints, however.

at any instant can be specified by 6 coordinates and might be represented by a "point" in 6-dimensional space. In fact, physicists are usually interested not only in position but in velocity as well. Now in the above example the velocity of each ball at any instant can be specified by two coordinates (component velocities). The position and velocities of the system together are called the "phase" of the system. In the above example the "phase" of the system at any instant may be represented as a point in 12-dimensional space, that is, as a set of 12 real numbers. Physicists call this 12-dimensional space the phase-space of the dynamical system. The totality of phases for all instants forms some kind of a locus in the 12-dimensional phase-space and the physical properties of the dynamical system may be studied by means of the geometric properties of this locus. This is actually done in advanced physics.

Another example of the use of many dimensions in physics occurs in the Special Theory of Relativity. This example has needlessly bewildered many people. In this subject, the things being studied are "events," such as the flashing of a light signal at a given time and place. The word "event" is not used here in its everyday sense as something unusual; the mere existence of a particle of matter in a certain place at a certain time is an "event." Thus, an event may be specified by 4 real numbers: three of them, x, y, and z, say, are coordinates of the place where the event occurs, and the fourth, t, say, tells the time of its occurrence. Hence an "event" may be represented as a "point" in a "space" of 4 dimensions. (In the later General Theory of Relativity the four coordinates lose their individual meanings, because of certain changes or transformations of the coordinate system, and the fourth can no longer be said to stand for time; but we cannot go into details here.) Let us hope that this brief explanation will dispel any aura of mystery surrounding the Fourth Dimension (always spelled with capitals to make it seem mysterious) about which many people, who have had the courage of their confusion, have written a great deal of nonsense. If you observe anyone talking or writing about the Fourth Dimension in an exalted, mystic tone you may bet with a fair degree of safety that he is not a mathematician but a Bohemian Philosopher, possibly with a concentrated solution of alcohol in his bloodstream. It is

also worth pointing out that brilliant new ideas lie at the foundation of the theory of Relativity but mere use of 4-dimensional space is not one of them. Mathematicians have studied the geometry of many-dimensional space since the middle of the 19th century and physicists have been using spaces of more than 4 dimensions, as phase-spaces, for example, about as long. The fact that 4 dimensions are convenient in Relativity is something of a triviality; the excellence of the theory is to be found elsewhere.

89. Constructions with ruler and compasses. Now that we know a little analytic geometry, we can indicate how the criterion for constructibility by ruler and compasses (which was discussed in section 59, Chapter VIII) may be established. This criterion says essentially, that *we can construct, with ruler and compasses alone, all those and only those quantities which can be obtained from the given quantities by a finite number of rational operations and extractions of square roots.* The "all those" part of this statement was established in section 59; there remains only the "only those" part of it to be discussed. Now, in constructing a figure with ruler and compasses one deals exclusively with the intersection of two lines, the intersections of a line with a circle, and the intersections of two circles. Finding the intersection of two lines amounts algebraically to solving simultaneously two linear equations in x and y. This (see section 84, example 1) is accomplished with the help of rational operations (addition, subtraction, multiplication, and division) alone (see section 39). Finding the intersections of a line with a circle amounts to solving simultaneously a linear and a quadratic equation in x and y. This (see section 84, example 2) reduces to the solution of a quadratic equation in one variable which, as you know, may be solved by a formula (section 36) which involves only rational operations and the extraction of a square root. Finding the intersections of two circles reduces to finding the intersections of one of the circles with their common chord (see section 84, example 3). This also involves only rational operations and square roots. This brief discussion is hardly a formal proof of our underlined statement, but it is not hard to see that these considerations might lead to a proof.

REFERENCES

E. A. Abbott, *Flatland, A Romance of Many Dimensions*, Little Brown, Boston, 1929.

E. T. Bell, *Men of Mathematics*, Simon and Schuster, N. Y., 1937 (biographies of Descartes, Fermat, and others.)

A. Dresden, *An Invitation to Mathematics*, Henry Holt, N. Y., 1936.

A. S. Eddington, *The Nature of the Physical World*, Macmillan, N. Y., 1928.

A. S. Eddington, *Space, Time, and Gravitation*, Cambridge University Press, Cambridge, 1920.

A. Einstein and L. Infeld, *The Evolution of Modern Physics*, Simon & Schuster, N. Y., 1938.

F. L. Griffin, *An Introduction to Mathematical Analysis*, Houghton Mifflin, Boston, 1921.

E. Kasner and J. Newman, *Mathematics and the Imagination*, Simon & Schuster, N. Y., 1940.

H. P. Manning, *The Fourth Dimension Simply Explained*, Munn and Co., N. Y., 1910.

Osgood and Graustein, *Plane and Solid Analytic Geometry*, Macmillan, N. Y., 1921.

B. Russell, *The ABC of Relativity*, Harper & Bros., N. Y., 1925.

A. N. Whitehead, *Introduction to Mathematics*, Henry Holt, N. Y., 1911.

Histories of Mathematics, as before.

Chapter X

FUNCTIONS

90. Variables and functions. By a **variable** we mean a symbol which may represent (many) different objects during the same discussion. By a **constant** we mean a symbol which may represent only one definite object during the same discussion. It is customary to use early letters of the alphabet like a, b, c for constants and later letters like x, y, z for variables. The set of all objects which a variable is permitted to represent during a discussion is called the **range** of the variable. Any object which the variable may represent (that is, any object in its range) is called a **value** of the variable. The "objects" in the range will usually be numbers. Note that a and b stand for constants even though their values are not stated. Of course, definite numbers like 2 or 3/2 are constants.

Consider two variables, x and y, each with its own range. The variable y is said to be a **function of x** when any scheme (rule, relation, or correspondence) is given whereby to each value of x there correspond one or more definite values of y. The scheme or rule whereby the value (or values) of y, corresponding to a given value of x, is determined, is called the **functional relation.** The variable x is called the **independent variable** and y is called the **dependent variable,** because the value (or values) of y is determined when a value of x is chosen. If the scheme associates to each value of x just one value of y, the function is called **single-valued,** otherwise, **many-valued.** The function is said to be **defined** for each value of the independent variable in its range. The significance of these definitions will become clear as soon as we examine some concrete examples.

Remark. The word function is used here in a technical sense quite different from its everyday meaning as in the statements "digestion of starch is a function of the saliva," or "passing bills

is a function of the legislature," or "the Senior Prom was the outstanding function of the school year."

Example 1. Let the range of x be the set of all integers. Then the function y defined by the relation $y = 2x + 1$ associates with each value of x just one value of y. The range of y is the set of odd integers. For example, to the value $x = 3$, it associates the value $y = 7$; to the value $x = 2$, it associates the value $y = 5$. The function is single-valued since to each integer x there corresponds just one odd integer y. We might have allowed the range of x to be the set of all rational numbers, or all real numbers, or all complex numbers, at the same time changing the range of y accordingly, of course.

Example 2. Consider the function y defined by the relation $y = \dfrac{1}{x - 3}$. Whatever set of numbers we choose for the range of x we must not include the number 3; for $x = 3$ this function is not defined. (Why?) We may let the range of x be the set of all real numbers except 3. The function is single-valued.

Example 3. Consider the function y defined by the relation $y^2 = x$ or $y = \pm\sqrt{x}$. This is a many-valued function of x, or more precisely a "two-valued" function of x, since for each value of x except 0 there will be two corresponding values of y. For example, if $x = 9$, $y = \pm 3$. We have assumed that the range of x is some set of numbers. If we want the range of y to contain only real numbers we must restrict the range of x to the set of non-negative real numbers; for if x were permitted to have a negative value, y would be imaginary. If we want the range of y to be restricted to integers, we must restrict the range of x to the set of integers which are squares of integers.

Example 4. Let the range of x be the set of all objects for sale in a certain department store. Let the value of y corresponding to each value of x (that is, each object) be the price of x. This functional relation defines y as a single-valued function of x.

Example 5. Let the range of x be the set of husbands in a given country. Let y be the "wife of x." This relation defines y as a function of x. It is a single-valued function if the range of x is the set of law-abiding husbands in a monogamous civilization.

Otherwise it may be a many-valued function. Notice that this function would still be single-valued in polyandrous civilizations but not in polygamous civilizations.

Example 6. Let the range of x be the set of all positive numbers with three consecutive significant digits; for example, 278 or .00278. Then the table of logarithms, properly read, associates with each such number x a number called its logarithm. Thus $y = \log x$ defines y as a single-valued function of x. Actually the range of this function can be taken as the set of all positive real numbers. But our table tells us the (approximate) value of $\log x$ only for the range stated above.

Example 7. Let the range of x be the set of years between 1930 and 1936. If y is the number of 4th of July accidents in the year x, then the correspondence given by the following (fictitious) table

x	1930	1931	1932	1933	1934	1935	1936
y	420	350	358	200	250	262	276

defines y as a single-valued function of x for the given range of x.

Example 8. Let x be the age of a certain baby, measured in days, and let y be its weight, measured in pounds, at 5 P.M. of the corresponding day. Consider the table:

x	0	7	14	21	28	35	42	49
y	$7\frac{13}{16}$	$8\frac{4}{16}$	$8\frac{9}{16}$	$9\frac{2}{16}$	$9\frac{9}{16}$	$10\frac{1}{16}$	$10\frac{9}{16}$	11

Here y is a single-valued function of x, for the given range of x.

Example 9. Let x be the number of days a patient has been in the hospital for a certain illness. Let y be the temperature of the patient, taken at noon, for the corresponding day. Consider the table:

x	1	2	3	4	5	6	7	8
y	102	103	103	101	100	99	98.6	98.6

Here y is a single-valued function of x, for the given range of x.

Example 10. Let the range of x be the set of all places in the United States. Let the range of y be the set of all points on a map of the United States. Let "y = the point on the map corresponding to the place x" be the functional relation. Then y is a single-valued function of x, since to each place x there corresponds just one point y on the map.

Example 11. Let the range of x be the set of all normal human beings. Let the range of y be the set of all feet. The functional relation "y is the right foot of x" determines y as a single-valued function of x. The functional relation "y is a foot of x" determines y as a double-valued function of x.

The extremely general idea of a function given here is a modern one. It may seem too general to be of much practical value. However, even these very general functions can be studied to good advantage. In elementary mathematics and for many applications *we can and will restrict ourselves to functions where the ranges of both variables are confined to real numbers.* In fact, in this book, we shall usually restrict ourselves still further to the consideration of *single-valued functions.* For example, the relations $y = \log x$ (where the range of x is the set of positive real numbers), or y = any rational expression in x define single-valued functions.

EXERCISES

In each of the following 6 exercises

(*a*) *find the largest possible set of real numbers which can be taken as the range of x if the range of y is to be confined to real numbers;*

(*b*) *plot the graph of each function:*

1. $y = 2x + 1$. **2.** $y = x^2$. **3.** $y = \pm\sqrt{x}$. **4.** $y = 1/x$.

5. $y = \pm\sqrt{9 - x^2}$. **6.** $y = \pm\sqrt{x^2 - 9}$.

7. Give three examples each of (*a*) variables, (*b*) constants, (*c*) functions, which occur in any connection or subject whatever.

91. Functional notation. If we have occasion to mention a certain function of x, such as $\dfrac{12 - x}{x}$, very often, we find it convenient to denote it briefly by a symbol like $f(x)$, read "f of x." Note that $f(x)$ does not mean a number f multiplied by a number

x, but stands for a "function of x"; in the given example, $f(x) = \dfrac{12 - x}{x}$. If a is any value of x (that is, any number in the range of x) then $f(a)$ *shall mean the value of the dependent variable* (*that is, the value of the function*) *corresponding to the value* $x = a$ *of the independent variable. That is,* $f(a)$ *is the value of* $f(x)$ *when* x *is replaced by* a. In the above example, $f(3) = \dfrac{12 - 3}{3} = 3$, $f(4) = \dfrac{12 - 4}{4} = 2$, $f(2) = \dfrac{12 - 2}{2} = 5$, $f(1) = \dfrac{12 - 1}{1} = 11$, $f(12) = \dfrac{12 - 12}{12} = 0$, and $f(0)$ does not exist, since 0 cannot be in the range of x for this function. (Why?) If we have to discuss another function of x at the same time, we may denote the second function by $F(x)$, $f_1(x)$, $g(x)$, $h(x)$, $h'(x)$, etc. For example, if $g(x) = 2x + 1$ then $g(1) = 3$, $g(2) = 5$. Hence, $f(3) = g(1)$ and $f(2) = g(2)$ in our examples. This functional notation is very compact and convenient.

EXERCISES

1. If $f(x) = 2x^2 - 3x + 4$, find: (*a*) $f(0)$. (*b*) $f(1)$. (*c*) $f(2)$. (*d*) $f(-2)$. (*e*) $f\left(\frac{1}{2}\right)$.

2. If $f(x) = 8^x$, find: (*a*) $f(1)$. (*b*) $f(2)$. (*c*) $f(0)$. (*d*) $f(1/3)$. (*e*) $f(-2/3)$. (*f*) $f(-2)$.

3. If $g(x) = \dfrac{1}{x-2} + \dfrac{3}{x-5}$, find: (*a*) $g(1)$. (*b*) $g(3)$. (*c*) $g(7)$. (*d*) $g(1/2)$. (*e*) What numbers must be excluded from the range of x?

4. (*a*) If $f(x) = x^2 + 1$ and $g(x) = 9 - x^2$ show that $f(2) = g(2)$ and $f(-2) = g(-2)$.

(*b*) Plot the graphs of both functions on the same coordinate system and interpret the latter statements graphically.

(*c*) Is $f(-2) = -f(2)$?

5. If $f(x) = 3x^2 + 2x + 3$, find: (*a*) $f(2)$. (*b*) $f(a)$. (*c*) $f(a + h)$. (*d*) $f(a + h) - f(a)$. (*e*) $\dfrac{f(a + h) - f(a)}{h}$.

Find the value of $\dfrac{f(x + h) - f(x)}{h}$ *for each of the following functions:*

6. $f(x) = x^2$. **7.** $f(x) = x^3$. **8.** $f(x) = 4x^2 + 3x + 1$.

9. $f(x) = 1/x$. **10.** $f(x) = 3 - 2x - 3x^2$. **11.** $f(x) = 1/x^2$.

92. The graph of a function. The study of functions began to be very important in the 17th century when scientists became greatly interested in the study of varying quantities. The invention of the pendulum clock, the problems of transoceanic navigation, the study of the motions of cannon balls and planets, and other physical phenomena, all contributed to the need for this study. If y is a function of x, we say that the value of y *depends* on the value of x. Speaking loosely, we might say that y *varies with* x*;* or a change in x *induces* a corresponding change in y. This statement is not strictly true for we might have a function which assigns the same value to y no matter what value is given to x. Such a function might be called a *constant function;* its graph would be a horizontal straight line. But, it will do no great harm if you think of the value of y changing when the value of x changes provided you agree that the change in y may be zero.

Examples of functions important in physical science are extremely numerous. Thus the distance through which a body falls at a given place on the earth's surface depends on the number of seconds during which it falls. The length of a steel bar depends on its temperature. The time required for a complete oscillation of a pendulum depends on the length of the pendulum. The distance through which a cannon ball, fired with a given initial velocity, will travel depends on the angle at which it is fired; the area of a square depends on the length of its side; and so on. However, to know merely that one quantity, y, depends on another, x, is hardly enough. It is usually essential to know whether y increases or decreases as x increases, and whether these changes are rapid or slow. More precisely, we might want to know how much y increases or decreases when x increases a certain amount; or how fast y is changing near a given value of x. We might want to know whether or not y has maximum or minimum values, and, if so, what they are; and so on. Some of these questions will be answered by a careful study of functions in Chapter XI.

For the present, we can see that a rough idea of how a function $y = f(x)$ varies may be obtained from its graph. Graphical methods are used extensively in many fields, especially where approximate results are desired.

Example 1. Consider the function $y = 10 - x^2$. To draw its

graph we first choose a number of values of x and calculate the corresponding values of $y = f(x)$; these results may be put into a table:

x	0	1	-1	2	-2	3	-3	4	-4
$y = f(x)$	10	9	9	6	6	1	1	-6	-6

Plotting these points (x,y), we obtain several points of the graph. If we have enough points to give us a fairly accurate idea of the shape of the curve, we join them by a "smooth" curve (Fig. 111). We may now use the graph to estimate values of y corresponding to values of x not in the table. Thus for $x = 1.5$ we might estimate $y = 8$ from the graph (Fig. 111); this is not correct, but it is not far wrong. From the actual formula for the function we obtain for $x = 1.5$ the value $y = 10 - (1.5)^2 = 7.75$. Naturally, the more points we actually plot and the more carefully we plot them, the more accurate our graphical estimates will be. We could judge from the graph that 10 is the maximum value of y and that there is no minimum value.

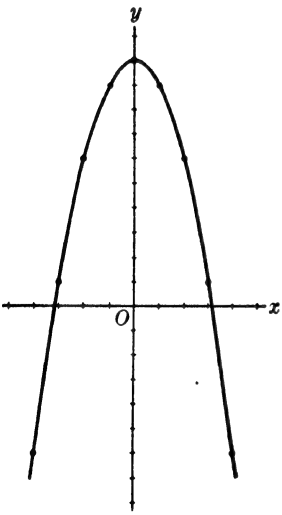


FIG. 111

When we join our plotted points by a smooth curve, we really assume a great deal. For instance, we assume that the curve does not behave queerly between the plotted points as in Fig. 112. In fact, we cannot be sure of this without a much deeper study of the function than we can make here. It is certainly clear that if we plot points corresponding to values of x which are too far apart, we may get an erroneous idea of the shape of the curve. But it is not always obvious how close together the values of x must be in order to avoid being "too far apart";

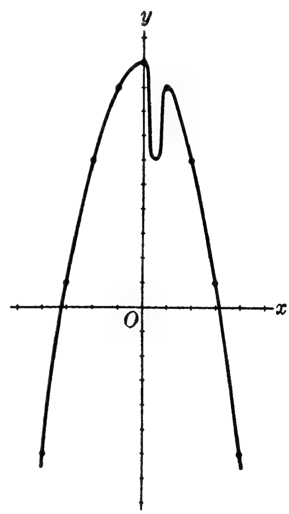


FIG. 112

consider the following examples. *Note that it is often advisable to use units of different lengths on the two axes in order to confine the graph to a moderate space.*

Example 2. Let the function be $y = 4x^3 - x$. As usual, we obtain a table:

x	0	1	-1	2	-2	$\cdots$
y	0	3	-3	30	-30	$\cdots$

If we plotted these points we might be tempted to join them as in Fig. 113. This would give us a *false* impression of the shape of

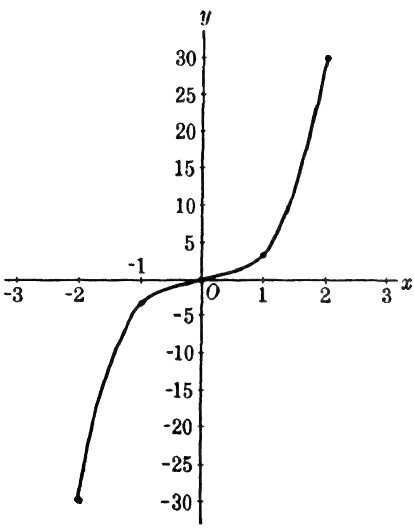


FIG. 113

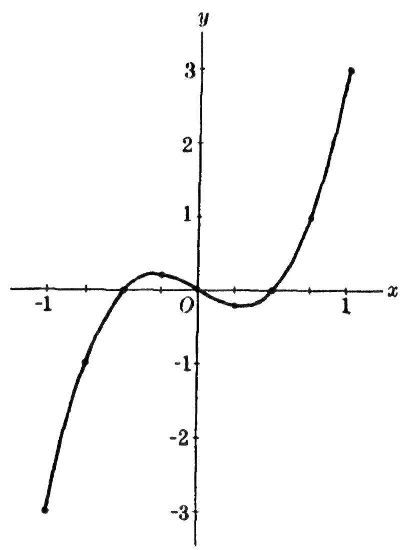


FIG. 114

the curve as we can see by using values of x chosen closer together, as in the following table, from which we see that the part of the graph between $x = -1$ and $x = 1$ (drawn on a different scale, for convenience) should look like Fig. 114:

x	0	1/4	$-1/4$	1/2	$-1/2$	3/4	$-3/4$	1	-1	3/2	$-3/2$	2	-2
y	0	$-3/16$	3/16	0	0	15/16	$-15/16$	3	-3	12	-12	30	-30

Example 3. Let the function be $y = 1/(2x - 3)$. We might obtain the following table:

x	0	1	-1	2	-2	3	4	...
y	$-1/3$	-1	$-1/5$	1	$-1/7$	$1/3$	$1/5$	...

Plotting these points we might suppose that the curve looked like Fig. 115. But, plotting more carefully, we see that for $x = 3/2$ there is no point of the curve. A more complete table would be:

x	0	1	5/4	11/8	23/16	...	3/2	...	25/16	13/8	7/4	2	3
y	$-1/3$	-1	-2	-4	-8	...	no value	...	8	4	2	1	1/3

This gives rise to the curve in Fig. 116. The curve is a hyperbola and has two separate branches.

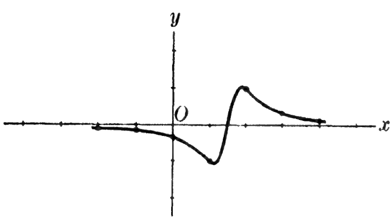


FIG. 115

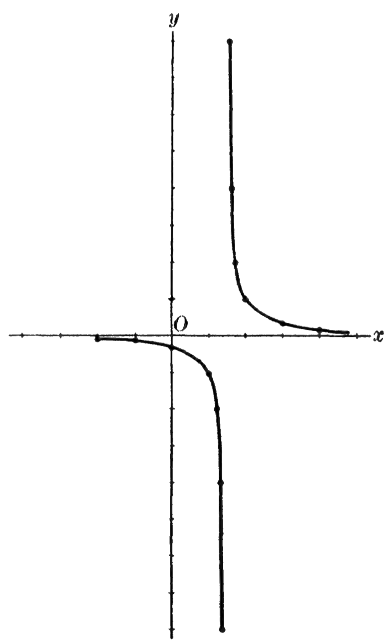


FIG. 116

Often we draw graphs for functions which are given originally in the form of a table. For example let us plot a graph for example 9 of section 90. Having plotted the points given in the table, we usually join them by a curve or a broken line (Fig. 117). If we want to estimate the temperature of the patient at midnight of the fifth day, we might read from the graph that it was approximately 100.5 degrees. But this might be wrong for the patient's temperature might have fluctuated considerably between noon of the 4th

day and noon of the fifth. However, if we plotted a graph for example 8, section 90, it would be more reliable to estimate the weight between the actually plotted points because we do not expect a baby's weight to undergo abrupt fluctuations. Thus, *great care must be exercised in reading from a graph more than has been put into it.* It is even possible to get complete nonsense if one is not careful. For example, a graph for example 7, section 90, is given in Fig. 118. If we were to use the graph to estimate values of the function between the plotted points, we might conclude that there were approximately 385 fourth of July accidents at Christmas 1930. In this example, joining the points by a line serves no purpose beyond that of a purely visual aid in seeing at a glance the trend of the function.

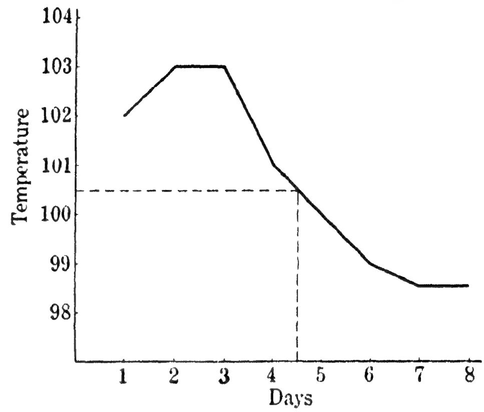


Fig. 117

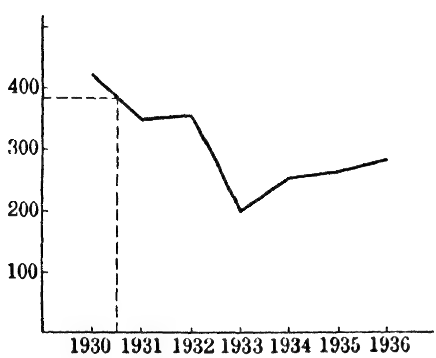


Fig. 118

In most of our examples the function was defined or given by means of a definite rule, or formula, or by a table, and the graph was derived from the definition of the function. However, a function may be defined by means of a graph directly and an approximate table of values, or an approximate formula may be derived from the graph. This actually occurs in various practical situations. Thus, a seismograph records on a strip of moving tape the graph of a function representing the vibrations of an earthquake. Similarly a cardiograph records on a strip of paper the graph of a function representing the vibrations of the heart beat. In the same manner, an instrument called a phonodeik records a graph of the vibrations of the air produced by a sound

wave. A phonograph record is essentially a graph of the sound scratched in wax so that a needle retracing the scratch (or graph) will reproduce the original vibrations and sound wave. In all these cases, the function is given originally as a graph and conclusions are drawn from the graph. For most purposes, a definite formula is more desirable than either a table or a graph since from it one can obtain as exactly as one pleases the value of y corresponding to any value of x.

Remark. Not all functions can be conveniently graphed. Thus consider the function defined as follows: let $f(x) = 1$ if x is a rational number and let $f(x) = 2$ if x is an irrational number. This function $y = f(x)$ is defined for all real values of x; that is, the range of x is the set of all real numbers. But between every pair of rational numbers (no matter how close together) there is an irrational number; and between every pair of irrational numbers (no matter how close together) there is a rational number. Hence between every pair of values of x for which $f(x) = 1$ there is a value of x for which $f(x) = 2$; and between every pair of values of x for which $f(x) = 2$ there is a value of x for which $f(x) = 1$. A graph cannot be drawn which will give a clear impression of the way in which this function varies. Such "pathological" functions as this are studied in advanced mathematics but are of no importance in elementary applications to physical science, for example, where all functions are supposed to vary in a "smooth" or "continuous" manner, instead of jumping fitfully between two values as this function does. Thus if a train is travelling from Washington to New York, we do not expect it to move gradually from Washington to Philadelphia and then suddenly disappear from Philadelphia and pop up in New York without any lapse of time and without traversing the intervening territory. We expect it to move continuously over all the places between Philadelphia and New York.

EXERCISES

1. (*a*) Plot the graphs of each of the following functions between $x = -5$ and $x = 5$, if possible;

(*b*) estimate from the graph the values of x for which $y = 1$; $y = 3/2$; $y = 0$;

(*c*) describe how y changes as x goes from -5 to $+5$. For what values of x does y appear to be a maximum? minimum?

(*i*) $y = 10 - \frac{1}{2}x^2$. (*ii*) $y = 3 - 2x$. (*iii*) $y = x^3 - 9x^2 + 24x - 7$.

(*iv*) $y = 8^x$. (Hint: plot values of x at intervals of 1/3.)

(*v*) $y = \log x$. (Hint: use table of logarithms.)

(*vi*) $y = 10^x$. (Hint: use table of logarithms.) (*vii*) $y = \frac{1}{x-2}$.

2. In the function $y = ax^2$, a being a constant different from zero, how does y change when x is doubled? tripled?

3. The distance s, measured in feet, through which a body falls in t seconds is found experimentally to be given approximately by the function

(1) $$s = 16\,t^2.$$

(*a*) Make a graph of this function, plotting integral values of t from $t = 0$ to $t = 4$.

(*b*) Estimate from the graph how high a cliff is if a body dropped from it reaches bottom in 1.5 seconds.

(*c*) If the cliff is 104 ft. high, in how many seconds will the body reach bottom?

4. (*a*) A bomb is dropped from an airplane 1600 ft. high. In how many seconds will it reach the earth? (Use formula (1).)

(*b*) If a bomb strikes the earth in 7 seconds, how high was the airplane when the bomb was dropped?

5. If a ball is thrown directly upward from an initial height of h_0 ft., with an initial velocity of v_0 ft. per sec., its height h, measured in feet, after t seconds is found to be given by the function

(2) $$h = h_0 + v_0t - 16\,t^2.$$

Assume $h_0 = 5$ ft., and $v_0 = 128$ ft. per sec.

(*a*) Make a graph of this function.

(*b*) After how many seconds will the ball strike the ground?

(*c*) After how many seconds will the ball reach its maximum height? What is the maximum height?

(*d*) After how many seconds will the ball be 85 ft. high?

6. Make a graph of the function in example 8, section 90. Estimate from the graph, the approximate weight of the baby at 5 P.M. of the 24th day; the 26th day.

7. A rectangular area is to be enclosed by 40 ft. of fence.

(*a*) express the area A as a function of the length x of one side of the rectangle.

(*b*) Make a graph of this function using a horizontal x-axis and a vertical A-axis.

(*c*) From the graph, estimate the value of x which will make A a maximum and estimate the maximum value of A.

8. An open box is to be made from a square piece of cardboard 12 inches on each side by cutting out equal squares from the corners and folding up the sides.

(*a*) Express the volume V of the box as a function of the length x of the edge of the cut-out square.

(*b*) Make a graph of this function using a vertical V-axis and a horizontal x-axis.

(*c*) From the graph, estimate the value of x which will make V a maximum, and estimate the maximum value of V.

9. A closed box with a square base is to have a volume of 64 cubic feet.

(*a*) Express the total area A of the top, bottom and 4 sides as a function of the length x of the base.

(*b*) Make a graph of the function, using a vertical A-axis and a horizontal x-axis.

(*c*) Estimate from the graph the value of x which will make A a minimum, and estimate the minimum value of A.

10. An open box with a square base is to be made from 400 sq. in. of material.

(*a*) Express the volume V as a function of the length x of the base.

(*b*) Make a graph of this function with a vertical V-axis.

(*c*) From the graph, estimate the value of x which will make V maximum, and estimate the maximum value of V.

11. For a box with square ends to be sent by parcel post, the sum of its length and girth (perimeter of a cross-section) must not exceed 84 inches. Suppose a box has for its sum of length and girth exactly 84 inches.

(*a*) Express the volume V of the box as a function of the length x of the edge of the square ends.

(*b*) Make a graph of this function, with a vertical V-axis.

(*c*) Estimate from the graph the value of x which will make V maximum, and estimate the maximum value of V.

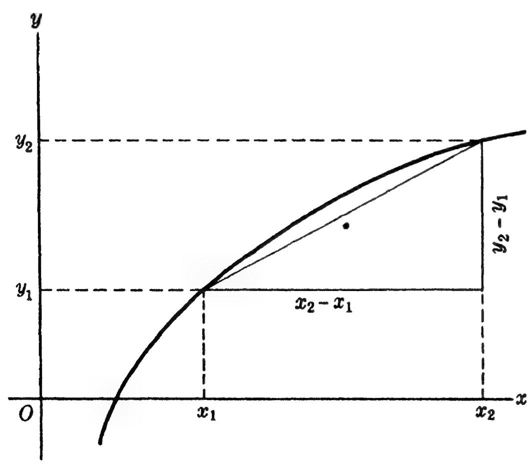


FIG. 119

93. Average rates. If an automobile travels 360 miles in 9 hours, we say that its average speed for the trip was 40 miles per hour. In general, *if y is a function of x, and y changes from y_1 to y_2 as x changes from x_1 to x_2, then* $\frac{y_2 - y_1}{x_2 - x_1}$ *is called* ***the average rate of change of y with respect to x for the interval between x_1 and x_2.*** That is, the average rate is simply the change in y divided by the change in x, or the change in y per unit change in x,

for the interval in question. If the graph of the function $y = f(x)$ happens to be a straight line, this average rate is clearly the slope of the line. If the graph is not a straight line, the average rate of change of y with respect to x for the interval between x_1 and x_2 is easily seen (Fig. 119) to be the slope of the chord joining the points (x_1, y_1) and (x_2, y_2). Thus average rates may be estimated from the graph.

Example. A falling body traverses s ft. in t seconds, where $s = 16\,t^2$ approximately. From the table

t	0	1	2	3	4
s	0	16	64	144	256

the average speed for the first second is 16 ft. per sec.; for the second second it is 48 ft. per sec.; for the first two seconds it is 32 ft. per sec.

EXERCISES

1. In the above example, find the average speed for (*a*) the third second; (*b*) the fourth second; (*c*) the first three seconds; (*d*) the first four seconds.

2. From the graph of example 1, section 92, find the average rate of change of y with respect to x for the interval between (*a*) $x = 1$ and $x = 2$; (*b*) $x = 1$ and $x = 3$; (*c*) What is the significance of the fact that these rates are negative?

3. From the graph of example 8, section 90, find the average rate of change of the weight per day for (*a*) the first 7 days; (*b*) the first 14 days; (*c*) the first three weeks.

4. From the graph of exercise 3, section 92, find the average rate of change for (*a*) the first second; (*b*) the first two seconds; (*c*) the interval between $t = 4$ and $t = 5$.

5. From the graph of exercise 5, section 92, find the average rate of change for (*a*) the first second; (*b*) the first two seconds; (*c*) the interval between $t = 2$ and $t = 4$.

94. Interpolation and extrapolation. The act of estimating values of a function y for values of x *between* those actually plotted or calculated is called **interpolation.** We have seen how we may interpolate roughly by reading from the graph of the function. Let us examine another method of interpolation that is used frequently, namely the method of **proportional parts.** We shall illustrate this method by the following example.

Example. Consider the function $y = 1 + 6x - x^2$, from which we get the following table:

x	0	1	2	3	4
y	1	6	9	10	9

For $x = 1.5$ we might obtain an estimate of the corresponding y as follows. As x goes from 1 to 2, y goes from 6 to 9; that is, y increases 3 units. Now $x = 1.5$ is halfway between 1 and 2; hence we might take a value of y halfway between 6 and 9, that is

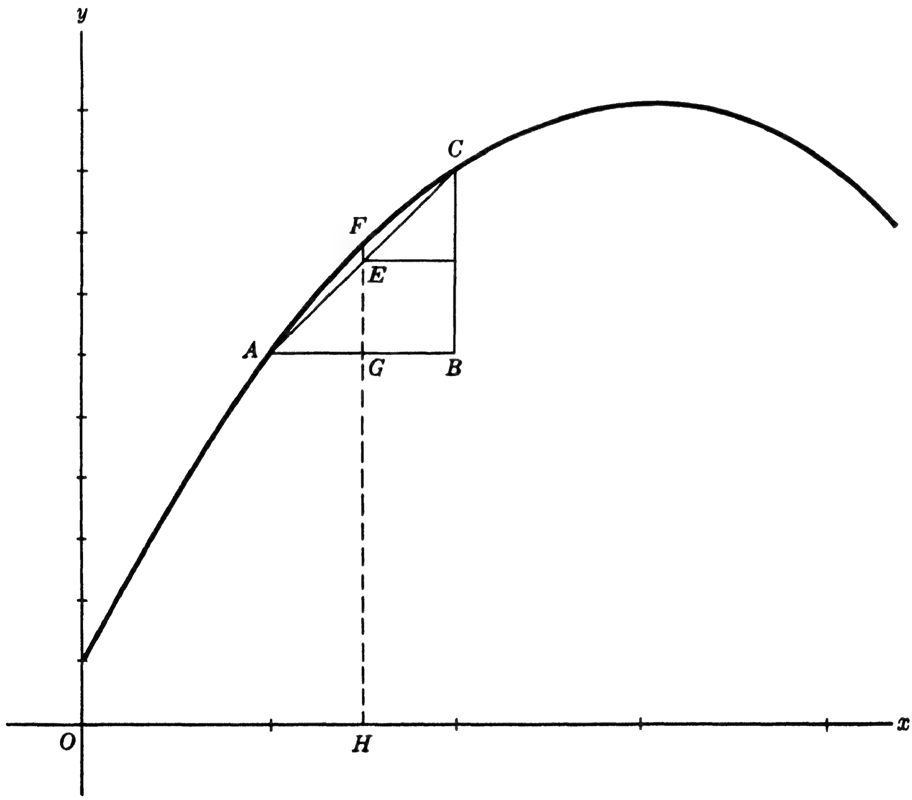


FIG. 120

$y = 7.5$, as our approximate value. Similarly for $x = 4/3$ which is 1/3 of the way between 1 and 2 we take for y a value 1/3 of the way between 6 and 9 or $y = 7$. These values can be seen to be inaccurate since we have a formula for the function which enables us to compute directly that for $x = 1.5$, $y = 7.75$ and for

$x = 4/3$, $y = 65/9 = 7\frac{2}{9}$; but they are not far wrong. Let us see why this is so. Draw the chord (straight line-segment) joining our two plotted points A and C (Fig. 120). For $x = 1.5$, the true value of y is the length of HF. But the midpoint E of the chord AC has the coordinates (1.5,7.5). This may be seen by means of the similar triangles AEG and ACB.

In general, *the method of interpolation by proportional parts amounts to using the ordinate of the point on the chord with the given abscissa instead of the ordinate of the point on the curve with the given abscissa. The accuracy of this method clearly depends on how closely the chord approximates the curve.* Thus if we interpolated by proportional parts for the value of y corresponding to the value $x = 1.5$ by using the points (0,1) and (3,10) we would get the result $y = 5.5$ which is far from the correct value $y = 7.75$. This is so because the chord joining the points (0,1) and (3,10) does not remain close to the curve. In general, the closer together are the values of x between which we interpolate, the closer will the chord approximate the curve. However, this statement cannot be relied on completely without further knowledge of the function or curve we are considering. For example, consider the function $y = 1/10x$. For $x = -.1$ we have $y = -1$ and for $x = .5$ we have $y = .2$. Suppose we were to interpolate by proportional parts to obtain an approximation for the value of y at $x = .2$, which is halfway between $-.1$ and $.5$. We obtain the value $y = -.4$ which is halfway between the values -1 and $.2$. But this is far from the true value $y = .5$. This happens because the chord is not at all close to the curve even though the values of x used are close together (Fig. 121). However, the method of interpolation by proportional parts can be shown to give a good approximation for most of the functions used in elementary mathematics and

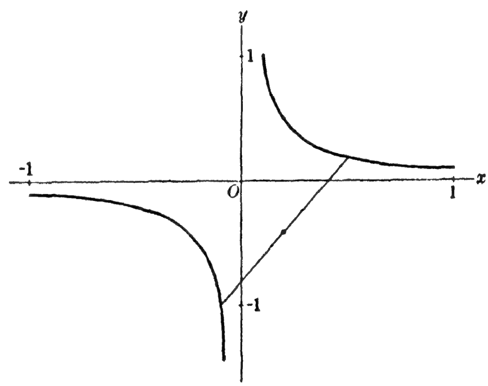


FIG. 121

physics provided we interpolate between values of x that are close enough together.

Because interpolation by proportional parts amounts to using the chord instead of the curve, it is often called **linear interpolation.**

The act of estimating values of a function *beyond* the range of values of x actually plotted or calculated is called **extrapolation.** Thus if we had plotted the curve of Fig. 120 only for $x = 0, 1, 2$, we might feel that since the curve is rising at $x = 2$, it will probably continue to rise as we proceed further to the right, at least for a short interval. Such a conclusion can often be justified, so far as the functions we meet in elementary mathematics and physics are concerned, but only for a *short* interval, as we see from Fig. 120 in this case. As for functions defined by a table of statistical observations, such extrapolation is extremely dangerous, even for a short interval.

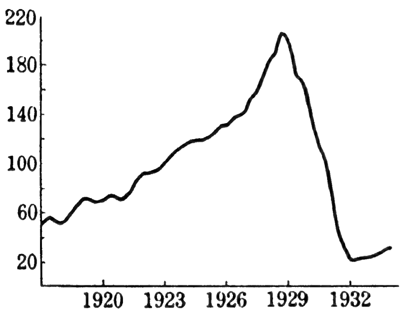


FIG. 122

Thus if the fictitious graph in Fig. 122 represented the average price of securities on the stock market from 1917 to 1932, we see how mistaken were the people who, in 1929, extrapolated incautiously and came to the conclusion that the rising trend would continue a while longer. To be at all secure in extrapolation, even more than in interpolation, one must know, by some means, the nature of the function one is dealing with. Thus in making predictions in physical science by extrapolation beyond our past and present observations in the case, say, of the return of a comet, or the occurrence of an eclipse, we are very safe, because we have a great deal of information as to the nature of the functions involved. But prediction by extrapolation from statistical data is not safe unless supported by sufficient additional knowledge of the functions involved.

EXERCISES

1. In the illustrative example $y = 1 + 6x - x^2$ above, interpolate by proportional parts from the given table for the value of y corresponding to (a) $x = 2.5$; (b) $x = 2.3$; (c) $x = 3.5$; (d) $x = .5$; (e) $x = .2$.

2. In example 8, section 90, interpolate by proportional parts to obtain the approximate weight of the baby at 5 P.M. on (*a*) the 11th day; (*b*) the 15th day. Extrapolate to obtain the approximate weight of the baby on (*c*) the 33rd day; (*d*) the 37th day; (*e*) Would we be justified in extrapolating to estimate the weight of the baby on the 400th day? its 20th birthday?

3. Interpolate to find (*a*) log 15.15; (*b*) log 15.12; (*c*) log 15.125; (*d*) log 15.175.

4. Interpolate to find the number whose logarithm is (*a*) 1.6744; (*b*) 1.6747; (*c*) 2.6742.

5. Interpolate to find the number whose logarithm is (*a*) 1.6924; (*b*) 2.6922; (*c*) 2.6926.

6. In how many years will any sum of money, invested at 6% interest compounded annually, double itself?

7. If $100 is invested at 4% interest compounded annually, how much will the amount be after 21 years?

8. If it is desired to have a sum of $1000 in the bank ten years from now and the bank pays 4% interest compounded annually, how much should be deposited now?

95. Classification of functions. An important chapter of pure and applied mathematics consists of the study of functions. To facilitate this study, it is convenient to classify functions in various ways.

One way is to distinguish between continuous and discontinuous functions. A function $y = f(x)$ may be called **continuous at the point (a,b)** if the dependent variable has the value $y = b$ corresponding to the value $x = a$ of the independent variable, and if the graph of the function consists of one connected piece in a small neighborhood of the point (a,b); otherwise it is called **discontinuous at $x = a$.** Thus the function $y = 1/x$ is discontinuous at the point where $x = 0$ but continuous for all other values of x. Most of the functions studied in elementary mathematics and its applications are continuous except possibly for isolated values of x. Polynomials are continuous for all values of x. In the next chapter we shall give a more precise definition of continuous function.

Example. The cost y of a taxi ride is a function of the distance x traversed. Suppose the cost is 20 cents for the first half mile and 10 cents for each succeeding half mile. Then the graph of this function is given in Fig. 123. Here y is a discontinuous func-

tion of x since the value of y jumps at the end of each half mile.*

Other modes of classification are often convenient. A function which can be expressed as a quotient of two polynomials in x is called a **rational function of x.** It can be shown that rational functions are continuous for all values of x except those for which the denominator becomes zero. Polynomials themselves are called **integral rational functions.** These terms are analogous to the terms rational number and integer. A polynomial of the first degree is called a **linear function.** If y is a function of x such that the values of y corresponding to every x can be obtained as roots of the same equation

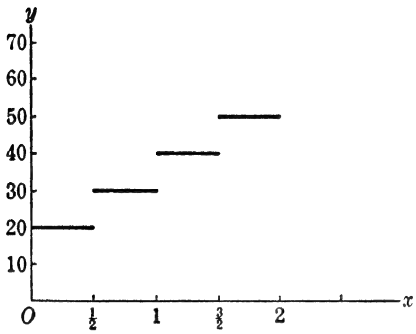


FIG. 123

$$f_0(x)y^n + f_1(x)y^{n-1} + \cdots + f_{n-1}(x)y + f_n(x) = 0$$

where the functions $f_0(x), f_1(x), \cdots, f_n(x)$ are polynomials in x then y is called an **algebraic function of x.** For example, $y = \sqrt{\dfrac{x^2+1}{x^2-1}}$ is an algebraic function of x since $(x^2-1)y^2 + (-x^2-1) = 0$; here $f_0(x) = x^2 - 1$, $f_1(x) = 0$, and $f_2(x) = -x^2 - 1$. All rational functions are algebraic for if $y = \dfrac{p(x)}{q(x)}$ where $p(x)$ and $q(x)$ are polynomials then $q(x)\cdot y - p(x) = 0$; here $f_0(x) = q(x)$ and $f_1(x) = -p(x)$. The functions $\log x$, 10^x, and the trigonometric functions which will be taken up in Chapter XII, can be shown to be not algebraic; they belong to a class of functions called **transcendental.** The terms algebraic function and transcendental function are analogous to the terms algebraic number and transcendental number introduced in section 62.

Functions like 10^x or c^x where c is any constant are called **exponential functions** because the independent variable x occurs as an exponent. They are important in many applications of mathematics, such as the following example.

Example. The number n of bacteria in a given culture at the

* Whether the value of this function y is 30 or 40 at the exact value $x = 1$ may depend on the size of the taxi driver (see Fig. 123).

end of t hours is given by $n = 100 \times 10^{.243t}$. Find ($a$) how many bacteria there are at $t = 0$; (b) at $t = 3$; (c) how many hours will be required for the number of bacteria to double itself.

(a) At $t = 0$ there are $n = 100 \times 10^0 = 100$ bacteria; (b) at t = 3 there are $n = 100 \times 10^{.729}$. Now $\log n = \log 100 + \log 10^{.729} = 2 + .729 \log 10 = 2.729$. Therefore, $n = 536.$, approximately. (c) n will be 200, as required, when t satisfies the equation $200 = 100 \times 10^{.243t}$. Hence $\log 200 = \log 100 + .243\, t \log 10$ or $2.301 = 2 + .243\, t$. Hence $t = \dfrac{.301}{.243} = 1.24$ hrs. approximately.

Special terminology is often used in physical science to refer to special types of functions which occur often. For example, if y is a linear function of x of the special form $y = cx$ where c is any constant whatever, then it is customary to say that **y varies directly as x.** The graph of this function is a straight line through the origin with slope c. If $y = c/x$ where c is any constant whatever, it is customary to say that **y varies inversely as x.** Thus Boyle's law for enclosed gases says that (at a given temperature) the pressure p varies inversely as the volume v; that is, $p = c/v$ where c is a constant. The graph of this function is a branch of a hyperbola, since only positive numbers are used. Newton's law of gravitation says that the gravitational force F exerted between any two particles whatever varies directly as the product of their masses m_1 and m_2 and inversely as the square of their distance d; that is, $F = Gm_1m_2/d^2$, where G is a constant. The constant G is known as the gravitational constant; when the units employed are centimeters, grams and seconds then $G = 6.66 \times 10^{-8}$ approximately. Note that if we ignore the decimal point (quite common practice among numerologists) then the gravitational constant is nothing more than the Number of the Beast (666) discussed in section 27. We leave this bit of information to any numerologist in good standing who wishes to make something of it.

Remark. The function $y = cx$ is a generalization of the multiplication table. For if c is a natural number, say 2, and if x takes only natural numbers as its values, the values of y are the results found in the "two times" table.

In the next chapter, we shall indicate some ways in which functions may be studied.

EXERCISES

1. If y varies directly as x and if $y = 21$ when $x = 7$,

(*a*) express y as a function of x.

(*b*) find the value of y when $x = 6$.

(*c*) how is y affected if x is tripled?

2. If y varies inversely as x and if $y = 4$ when $x = 3$,

(*a*) express y as a function of x.

(*b*) find the value of y when $x = 6$.

(*c*) how is y affected if x is doubled?

3. The maximum range of a projectile varies directly as the square of its initial velocity. If the maximum range is 15000 ft. when the initial velocity is 300 ft. per sec.:

(*a*) Write the range r as a function of the initial velocity v.

(*b*) What is the range when $v = 500$ ft. per sec.?

(*c*) What is v when $r = 60{,}000$ ft.?

4. Using Boyle's law above, assume that an enclosed gas kept at constant temperature has a volume v of 600 cu. in. when the pressure is 24 pounds per sq. in.

(*a*) Express p as a function of v.

(*b*) Find p when $v = 720$ cu. in.

(*c*) Find v when $p = 100$ lbs. per sq. in.

5. The weight of a body is essentially the gravitational force exerted on it by the earth. Hence, as a special case of Newton's law of gravitation above we deduce that the weight w of a body varies inversely as the square of its distance d from the center of the earth. Assuming this to be true and taking the radius of the earth as 4000 miles how much would a girl weigh at a height of 100 miles above the surface of the earth, if she weighs 100 lbs. on the surface of the earth.

6. Show that if y varies directly with x and if $y = y_1$ when $x = x_1$, and $y = y_2$ when $x = x_2$, then $\dfrac{y_1}{y_2} = \dfrac{x_1}{x_2}$. That is, corresponding values of x and y are in proportion.

7. Show that if y varies inversely as x and if $y = y_1$ when $x = x_1$, and $y = y_2$ when $x = x_2$, then $\dfrac{y_1}{y_2} = \dfrac{x_2}{x_1}$.

8. Using the same axes, draw the graphs of both $y = x^2$ and $y = 2^x$, between $x = 0$ and $x = 5$.

9. Using the same axes, draw the graphs of both $y = x^{-2}$ and $y = 2^{-x}$, between $x = 0$ and $x = 5$.

10. The number n of bacteria in a culture at the end of t hours is given by $n = 100 \times 10^{.0492\,t}$.

(*a*) Find the number present at the start ($t = 0$).

(b) How many were present at the end of 4 hours?
(c) In how many hours will the number present be 300?

11. A star loses heat in such a way that its temperature y, measured in degrees, at the end of x millions of years is given by $y = 15000 \times 10^{-.075x}$. Find its temperature at the end of 10 million years.

12. If the population P of a country, during a certain interval of time, is represented approximately by the formula $P = 4500000 \times 10^{.015t}$ where t is measured in years, find the population after 7 years.

13. If radium decomposes so that the number y of milligrams remaining at the end of t centuries is given by $y = 50 \times 10^{-.017t}$,
(a) How many milligrams were there at the start ($t = 0$)?
(b) How many milligrams were left at the end of 15 centuries?

96. Graphical solution of equations. Let $f(x)$ be any function of x. If the value of the function becomes zero when the number r is substituted for x, then r is said to be a **root** of the equation $f(x) = 0$. That is, by definition, a root of the equation $f(x) = 0$ is a number r such that $f(r) = 0$.* If r is a real number and r is a root of the equation $f(x) = 0$, then the graph of the function $y = f(x)$ must have a point in common with the x-axis at $x = r$. Therefore *the real roots of an equation $f(x) = 0$ are the x-coordinates of the points of intersection of the curve $y = f(x)$ with the x-axis* (Fig. 124). Consequently the real roots of an equation may be estimated graphically. Note that imaginary roots cannot be found in this way since on our graph x and y range only over real numbers.

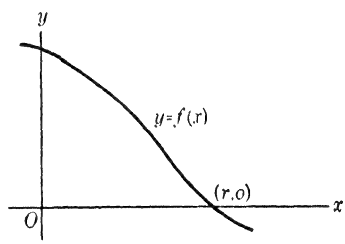


FIG. 124

Example 1. The roots of the equation $4x^3 - x = 0$ are $x = 0$, $x = 1/2$, and $x = -1/2$. See Fig. 114, page 269.

Example 2. Consider the equation $x^3 - x^2 - 5x + 5 = 0$. Plotting the graph of the function $y = x^3 - x^2 - 5x + 5$ for integral values of x we find the table:

x	0	1	2	3	-1	-2	-3
y	5	0	-1	8	8	3	-16

* Note that the statement $f(x) = 0$ is a propositional function since it contains the variable x, while the statement $f(r) = 0$ where r is a definite number or constant is a proposition.

From the graph (Fig. 125) we see that one root of the equation is 1, while there is another root between -2 and -3 and a third between 2 and 3. We could guess at the value of the root between 2 and 3, for example; but if we would like to get a closer estimate of this root we can locate it between successive tenths by plotting the graph (on a magnified scale, if desirable) for every tenth between 2 and 3, obtaining the table

x	2	2.1	2.2	2.3
y	-1	-0.649	-0.192	.377

This shows (Fig. 126) that the root of the equation between 2 and 3 is really between 2.2 and 2.3. If we wanted to approximate the root between successive hundredths we could now plot the curve between 2.2 and 2.3 (on a magnified scale, if desirable) using the values $x = 2.21, 2.22, 2.23, 2.24$ and so on until we discover between which two successive hundredths the curve crosses the x-axis; and so on.

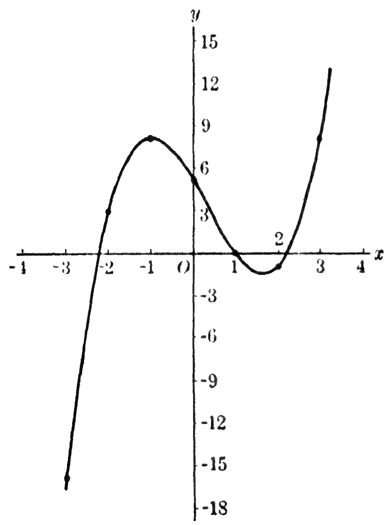


Fig. 125

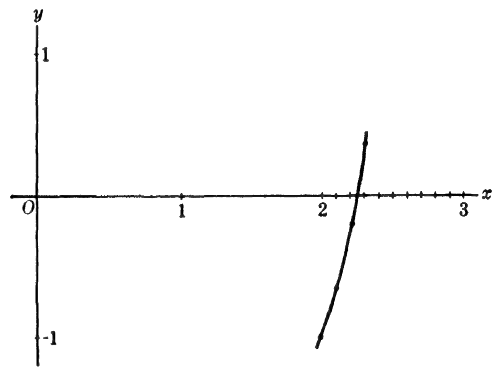


Fig. 126

The process of pinching or crowding the root between two successive units, successive tenths, successive hundredths, and so on, used in example 2, is a standard method of approximating the roots of any equation involving single-valued continuous functions. For if a single-valued function is continuous and its graph is below the x-axis at one value of x and above the x-axis at an-

other value of x, then the graph must cross the x-axis at least once somewhere between these two values of x. This method of hemming in the root, much as one runs down a base-runner caught between first and second base, is very useful although tedious. We have already used it in section 25 to approximate numbers like $\sqrt{2}$, $\sqrt[3]{2}$, etc. Note that $\sqrt{2}$ is a root of the equation $x^2 - 2 = 0$ and $\sqrt[3]{2}$ is a root of the equation $x^3 - 2 = 0$, and so on. Slight improvements in technique can be made which save some time but we shall not go into them here.

EXERCISES

Solve the following equations graphically:

1. $4x^2 - 8x + 3 = 0.$
2. $4x^3 - 13x + 6 = 0.$
3. $x^3 - 4x^2 + x + 6 = 0.$
4. $4x^3 - 24x^2 + 27x + 20 = 0.$
5. $2x^3 - 3x^2 - 9x + 10 = 0.$

Locate the real roots of the following equations between successive tenths:

6. $x^2 + 3x - 1 = 0.$
7. $x^3 - 3x^2 - 2x + 6 = 0.$
8. $x^3 - 4x + 2 = 0.$
9. $x^3 - 3x + 11 = 0.$
10. $x^4 - 4x - 5 = 0.$
11. $2^x + 3x - 6 = 0.$
12. $2^x - 8 + 4x = 0.$
13. $3^x + 5x - 10 = 0.$

97. Functions of several variables. It is frequently necessary to study functions of more than one independent variable. Thus the area of a rectangle is a function of its length x and width y, given by the formula $A = xy$. The perimeter of the rectangle is given by the formula $P = 2x + 2y$. The volume of a rectangular box is a function of three independent variables, the length x, width y, and height z, given by the formula $V = xyz$. The surface area of the box is given by $S = 2xy + 2xz + 2yz$. The amount of money A in a bank account on which interest is compounded annually is a function of 3 independent variables, the original principal P, the number n of years, and the rate r per cent at which interest is paid; this function is given by the formula $A = P(1 + \frac{r}{100})^n$. The study of functions of more than one independent variable presents technical difficulties very soon, and *we shall therefore confine ourselves, in our further study of functions, to functions of one independent variable.*

98. The application of functions in practical science. We have already indicated how widespread are the applications of func-

tions in science. We must again emphasize, however, the distinction between pure and applied mathematics. For example, the distance s, measured in feet, through which a falling body falls in t seconds is given approximately by $s = \frac{1}{2}gt^2$ where the constant g is found experimentally to be about 32. (Note: g is the acceleration, measured in feet per second per second, due to gravity.) Now we may use this formula to predict that a body will fall 64 feet in 2 seconds, 144 feet in 3 seconds, and so on. But even if our arithmetical calculations based on this formula are flawless, we must not be misled into believing that our results are perfectly accurate with respect to the physical facts. For the study of the *function* $s = 16\,t^2$ belongs to pure mathematics while the question whether or not the given function fits the physical facts accurately is a question of applied mathematics. In fact, careful experimentation would show that $g = 32.2$ would fit the facts more accurately than $g = 32$. Further experiment would show that the value $g = 32.1724$ is a still better approximation. In fact, the value g varies slightly depending on the latitude of the place on the earth where the body is falling and on the height above sea-level. Thus one should not confuse accuracy insofar as the purely logical perfection of pure mathematics is concerned with the accuracy with which the pure mathematics fits the concrete application.

 On the other hand, the accuracy of some of the formulas employed in physical science is amazing. The most familiar of these spectacular achievements are probably those of astronomy, wherein eclipses, etc., are predicted at long range. One of the most dramatic triumphs of applied mathematics was the discovery of the planet Neptune by Leverrier and Adams, about 1846. The planet Uranus had exhibited distressing unwillingness to move as the theory of gravitation said it should. It was suggested that the deviations might be due to the gravitational pull of some other body whose existence had not hitherto been suspected. Leverrier and Adams, independently of each other, performed the difficult mathematical task of calculating how large a body would have to be and how it would have to move in order to produce exactly the observed deviations in the orbit of Uranus. Leverrier wrote to an astronomer Galle telling him that this hy-

pothetical body should be found at a certain place at a certain time. Galle looked with his telescope and there it was! Thus the discovery of the planet Neptune, which had never been seen before that time, was made by purely mathematical prediction.

REFERENCES

E. T. Bell, *The Handmaiden of the Sciences*, Williams and Wilkins, Baltimore, 1937.

H. R. Cooley, D. Gans, M. Kline, H. E. Wahlert, *Introduction to Mathematics*, Houghton Mifflin, Boston, 1937. (pp. 245–432 on functions.)

M. E. J. Gheury de Bray, *Exponentials Made Easy*, Macmillan, London, 1921.

F. L. Griffin, *Introduction to Mathematical Analysis*, Houghton Mifflin, Boston, 1921.

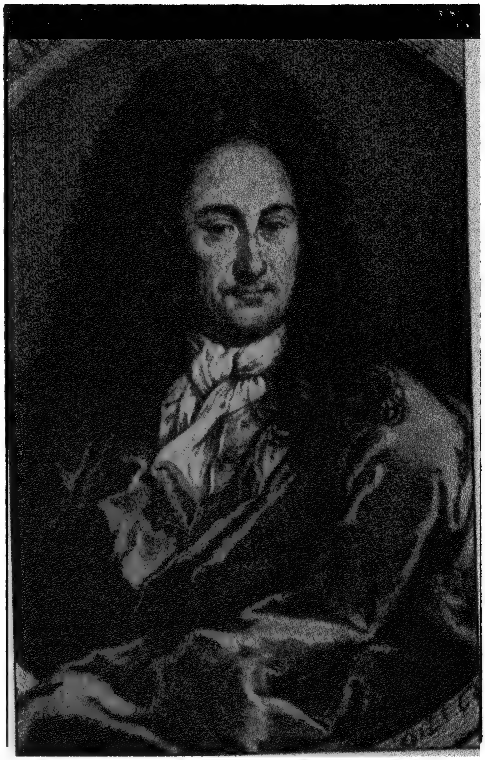
A. N. Whitehead, *Introduction to Mathematics*, Henry Holt, N. Y., 1911.

Chapter XI

LIMITS AND THE CALCULUS

99. Introduction. We have seen how important the study of varying quantities was in the physical science of the 17th century when the motions of planets, pendulums, cannon balls, sounds, etc., came to be of central interest. Both the physics of a changing world and the study of problems connected with tangents to curves, which arose in the analytic geometry of Descartes and Fermat, led to the invention of the calculus. The invention of the (differential and integral) calculus was accomplished by the mathematician and physicist Isaac Newton (1642–1727), most famous for his remarkable theory of gravitation, and Gottfried Wilhelm Leibniz (1646–1716), the mathematician and philosopher. Needless to say, neither Newton nor Leibniz created the calculus out of thin air. Other people, Fermat, for example, had some of the ideas of the calculus before these two, and some of the problems involved had been attacked with partial success years earlier. Newton himself is said to have remarked "If I have seen a little farther than others it is because I have stood on the shoulders of giants." The fact that progress is

seldom independent of the past is amply illustrated in the history of science. This in no way detracts from the gigantic contributions of Newton and Leibniz. They developed and organized the subject, making it one of the most powerful chapters of mathematics. The invention of the calculus initiated a period of exceedingly rapid development both in mathematics and in its applications to physics, astronomy, engineering, etc. In the latter half of the 17th century and the entire 18th century, much effort was devoted to the important task of developing further the ideas of Newton and Leibniz and their manifold applications in physical science.



Gottfried Wilhelm Leibniz
1646–1716, German 14

That the ideas of the calculus were "in the air" in the 17th century is partly indicated by the fact that Newton and Leibniz obtained their results independently of each other, although at the time each was accused of plagiarizing from the other. This controversy developed into one of the most heated feuds in the history of science (which is rich in disputes about priority) and helped to create an almost complete schism between British and German mathematicians. The unworthy kind of nationalism which displayed itself in this and other such controversies is merely evidence that mathematicians are also human beings often subject to the same weaknesses as other mortals. What is more inspiring, however, is the fact that the thread of scientific progress has persisted through the ages and has been contributed to by many different civilizations, nationalities, races, and creeds, and has come to be one of the most precious parts of man's heritage.

The entire calculus is based on the concept of function and the concept of "limit." The notion of "limit" was, in fact, familiar to the ancient Greek philosophers and mathematicians. Indeed, some of them had some of the essential ideas of the calculus. But they failed to develop any usable technique in connection with their ideas and hence made little progress. The concept of limit may be based on the notion of "sequence" which we shall take up now.

100. Sequences. By a **sequence,** (or an **infinite sequence**), of numbers we mean an unending succession of numbers x_1, x_2, x_3, $\cdots$, x_n, $\cdots$, where the dots mean "and so on." The subscript n merely indicates the position of the number x_n in the sequence. Thus x_1 is the first term, x_2 is the second term, x_3 is the third term. The term x_n is called the ***n*th term** or the **general term.** A sequence may be defined by stating its general term. For example, the sequence whose general term is $\frac{n}{n+1}$ is the sequence $1/2, 2/3, 3/4, 4/5, 5/6, \cdots, \frac{n}{n+1}, \cdots$. The sequence whose general term is $1 - \frac{1}{10^n}$ has for its first few terms the numbers $.9, .99, .999, .9999, .99999, .999999, \cdots$. The sequence whose general term is $\frac{(-1)^n}{n}$ has for its first few terms the numbers $-1/1, 1/2, -1/3, 1/4, -1/5, 1/6, \cdots$.

A sequence can be regarded as a single-valued function of n where the range of n is the set of all natural numbers; to each natural number n there corresponds a number x_n, the nth term in the sequence. Thus the symbol x_n might have been written as $x(n)$ to emphasize that it is a function of n.

Remark. Infinite sequences may arise outside of what is usually considered to be mathematics. For example, if a label on a can has on it a picture of the can, it gives rise to an infinite sequence of pictures of the can.

EXERCISES

Write the first five terms of the sequence whose general or nth term is:

1. $2 + \frac{1}{10^n}$. **2.** $1/2^n$. **3.** $\frac{2^n - 1}{2^n}$. **4.** $2 - \frac{1}{10^n}$.

5. $\dfrac{(-1)^n}{2^n}$. 6. $\dfrac{(-1)^{n+1}}{10^n}$. 7. $1/n$. 8. $3 + \dfrac{(-1)^n}{n}$.

9. $1 + \dfrac{1}{2^n}$. 10. $1 + \dfrac{(-1)^n}{2^n}$.

101. Limit of a sequence. Let us mark on a line the numbers of the sequence whose nth term is $2 - \dfrac{1}{10^n}$; that is, 1.9, 1.99, 1.999, 1.9999, $\cdots$. It is intuitively clear from Fig. 127 that as n

FIG. 127

increases indefinitely, the numbers $2 - \dfrac{1}{10^n}$ become closer and closer to 2; furthermore it can be shown that $2 - \dfrac{1}{10^n}$ becomes *as close as we please to 2 if n is taken great enough.* Therefore, we say that 2 is the *limit* of the sequence. The latter condition is important, for $2 - \dfrac{1}{10^n}$ also gets closer and closer to 3 as n increases indefinitely but it does not get as close as we please to 3, since for any value of n the number $2 - \dfrac{1}{10^n}$ is more than 1 unit away from 3. In other words as n becomes large, the distance between 2 and the numbers $2 - \dfrac{1}{10^n}$ becomes as small as we please. Now the words small and large have only comparative, not absolute, meaning. Thus .001 is small compared with 1000000 but large compared to .00000000001; a hundred miles is a large distance to a lame man who has to walk it, but an insignificantly small distance to an astronomer. The condition that the terms of the sequence get as close to 2 as you please can be formulated as a game, as follows: if you choose a "small number" h, we can find a natural number N such that all the terms after the Nth are closer to 2 than the distance h. Thus if you choose $h = .001$ we can take $N = 3$, and it is clear that every term after $2 - \dfrac{1}{10^3}$ is closer to 2

than the distance .001 (that is, lies between 2 − .001 and 2 + .001). If you choose $h = .00001$ the value $N = 3$ will no longer do, but we can take $N = 5$, say. That is, having chosen a "degree of closeness" h we can find a place in the sequence so that every term thereafter is closer to 2 than the preassigned amount h. This leads us to the formal definition of the limit of a sequence.

* DEFINITION. *The number a will be called the* ***limit of the sequence*** $x_1, x_2, x_3, \cdots, x_n, \cdots$ *provided that, given any positive number h, no matter how small, there exists a corresponding term of the sequence* x_N *such that every succeeding term* x_m $(m > N)$ *lies between the numbers* $a - h$ *and* $a + h$ *(that is, the distance* $|x_m - a|$ *of every* x_m*, with* $m > N$*, from a is less than the preassigned amount h).*

This definition may seem complicated at first because of the unfamiliar precise language in which it is couched, but the idea is really simple. Intuitively it means that *the numbers* x_n *get to be as close as we please to a as n increases indefinitely.* Let us see what it means in an example.

Example. Prove that the sequence $-1, 1/2, -1/3, 1/4, \cdots, (-1)^n/n, \cdots$ has the limit 0. To do this we must show that given any positive number h, there exists a term of the sequence $1/N$ such that every succeeding term $1/m (m > N)$ lies between $0 - h$ and $0 + h$. This is clearly so. We have only to choose $1/N$ smaller than h (Fig. 128). Thus if we choose $h = .001$ we have

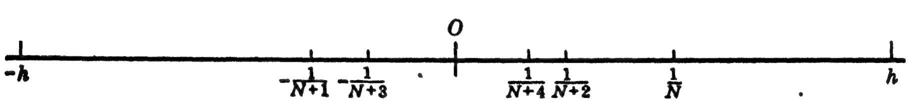


FIG. 128

only to take $N = 10{,}000$, since $+ 1/10000, - 1/10001, \cdots$ are all nearer to 0 than the distance .001. If we choose $h = .00001$ then $N = 10{,}000$ will not do, but we have only to take $N = 1{,}000{,}000$, say, since $1/1000000, - 1/1000001, \cdots$ are all nearer to 0 than the distance .00001.

If the sequence $x_1, x_2, x_3, \cdots, x_n, \cdots$ has the limit a, we write

$$x_1, x_2, x_3, \cdots, x_n, \cdots \rightarrow a \text{ (read, approaches } a\text{)}$$

* The formal definition may be omitted without disturbing the continuity of the chapter.

or $\quad$ as $n \uparrow$, $x_n \to a$ (read, as n increases indefinitely x_n approaches a)

or $\quad \lim\limits_{n\uparrow} x_n = a,$ (read, the limit of x_n as n increases indefinitely is a)

or simply $\quad \lim x_n = a,$ (read, the limit of the sequence x_n is a.)*

We say that x_n **tends to a** or **converges to a** or **approaches a as a limit as n increases indefinitely.** In this case we say the sequence is **convergent**; otherwise **divergent.** Thus the sequences

$$1,\ 2,\ 3,\ 4,\ \cdots,\ n,\ \cdots$$
$$2,\ 4,\ 6,\ 8,\ \cdots,\ 2n,\ \cdots$$
$$1,\ 4,\ 9,\ 16,\ \cdots,\ n^2,\ \cdots$$
$$-1,\ 1,\ -1,\ 1,\ \cdots,\ (-1)^n,\ \cdots$$
$$-1,\ 2,\ -3,\ 4,\ \cdots,\ (-1)^n n,\ \cdots$$

are not convergent since there is clearly no number to which the terms of these sequences get as close as we please. Note that we may have $\lim x_n = a$ even though none of the numbers x_n in the sequence is equal to a. That is, *the limit of a sequence need not be a term of the sequence.*

It is not always easy to prove that a sequence has a limit. However, in a more advanced course we would rest a great deal of work on this definition of limit. In fact, the entire subject of calculus may be based on it. Because of the technical difficulties, we shall here employ the idea of limit intuitively.

EXERCISES

1–10. What does the limit of each of the ten sequences given in the exercises in section 100 appear to be? What does this mean in terms of the definition of limit?

Make up a sequence whose limit is:

11. Five. **12.** Three. **13.** Zero. **14.** Ten.

102. Limit of a function.

Consider a sequence of values of x approaching a as a limit: $x_1,\ x_2,\ x_3,\ \cdots,\ x_n,\ \cdots \to a$. Suppose

* The notation $\lim\limits_{n\to\infty} x_n = a$ (read, the limit of x_n as n becomes infinite is a) is also widely used, but it often misleads the unwary student into believing that there is a peculiar number, called "infinity" and denoted by a lazy eight ∞, which n approaches. This is, of course, not the case.

that every term x_n of the sequence is in the range of the independent variable x for a given function $f(x)$; that is, $f(x_n)$ is defined. Then it may be that the sequence $f(x_1), f(x_2), \cdots, f(x_n), \cdots$ converges to a limit L. If $x_1', x_2', x_3', \cdots x_n', \cdots$ is another sequence of values of x converging to a, it might be that the sequence $f(x_1'), f(x_2'), \cdots, f(x_n'), \cdots$ does not converge at all, or possibly converges to a limit $L' \neq L$. These things happen in the functions whose graphs are given in Figs. 129 and 130, respec-

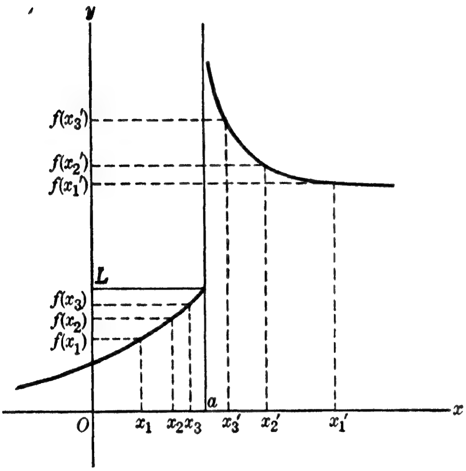


FIG. 129

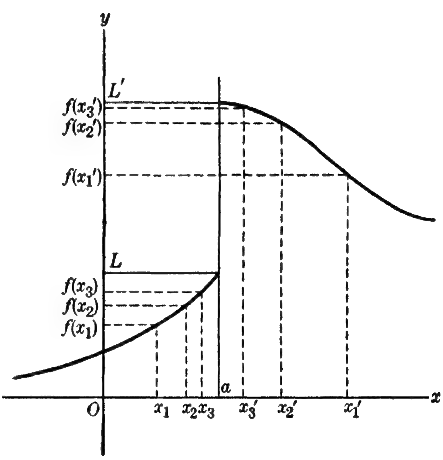


FIG. 130

tively. But it *may* happen that, no matter what sequence of x's we take converging to a, the corresponding values of the function will converge to the limit L. In this case, we say that $f(x)$ approaches the limit L as x approaches a. Precisely, we make the following definition.*

DEFINITION. *If for* **every** *sequence $x_1, x_2, x_3, \cdots, x_n, \cdots \to a$ (where each $x_n \neq a$), all x_n's being in the range of the independent variable x, we have $f(x_1), f(x_2), \cdots, f(x_n), \cdots \to L$, then we say that L is the* **limit of f(x) as x approaches a.** *In symbols we write $\lim_{x \to a} f(x) = L$, or $f(x) \to L$ as $x \to a$.*

Intuitively, the statement $\lim_{x \to a} f(x) = L$ means that *the value of $f(x)$ gets close to L as the value of x gets close to a.* One reason for the parenthetical remark in the definition is that we do not re-

* The formal definition may be omitted without disturbing the continuity of the chapter.

quire that the number a be in the range of the independent variable at all; that is, $f(a)$ need not be defined. We want to discuss what happens as we approach a even if we cannot be at a. This is illustrated in the following example.

Example 1. Let $f(x) = \dfrac{x^2 - 9}{x - 3}$. This function is defined for all values of x except $x = 3$. (Why?) For any $x \neq 3$, in fact, the value of $f(x)$ is the same as for the (different) function $y = x + 3$. Thus the graph of $y = \dfrac{x^2 - 9}{x - 3}$ is the same as the straight line $y = x + 3$ except that the point (3,6) is removed, since our function is not defined for $x = 3$. Nevertheless, $\lim\limits_{x \to 3} \dfrac{x^2 - 9}{x - 3}$ exists and is 6. That is, as x approaches 3, $f(x)$ becomes as close as we please to 6. Thus we see that $\lim\limits_{x \to a} f(x)$ can exist even when $f(a)$ does not.

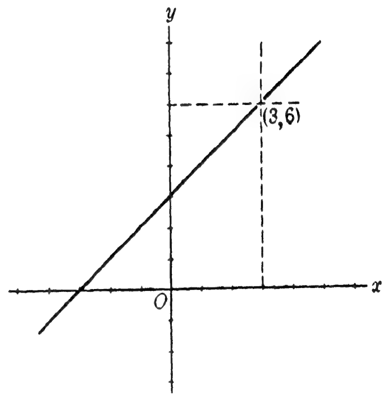


FIG. 131

Notice that $\lim\limits_{x \to 3} (x^2 - 9) = 0$ and $\lim\limits_{x \to 3} (x - 3) = 0$; yet,

$$\lim_{x \to 3} \frac{x^2 - 9}{x - 3} = 6.$$

Example 2. Consider the function defined as follows:

$$f(x) = \begin{cases} \dfrac{x^2 - 9}{x - 3} & \text{for all } x \neq 3 \\ 1 & \text{for } \quad x = 3. \end{cases}$$

This function is defined for *all* values of x. Its graph is given in Fig. 132. The point (3,6) is not in the graph while the isolated point (3,1) is in the graph. For this function we have $f(3) = 1$

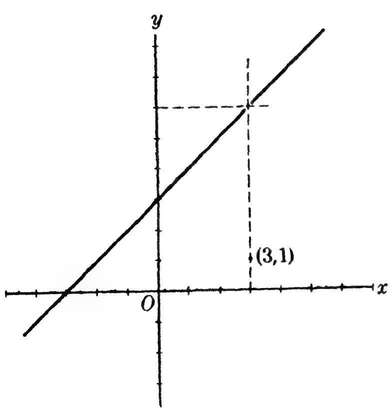


FIG. 132

while $\lim_{x \to 3} f(x) = 6$. This situation obviously corresponds to the "jump" in the graph at $x = 3$.

These are examples of discontinuous functions. They suggest the following definition.

* DEFINITION. *A single-valued function $f(x)$ is said to be* **continuous at x = a** *if:* (1) *$f(a)$ is defined;* (2) *$\lim_{x \to a} f(x)$ exists;* (3) *$\lim_{x \to a} f(x) = f(a)$. If any of these three conditions fails to be satisfied the function is called* **discontinuous at x = a.**

The third condition says that the limit of the function as x approaches a is the same as the value of the function at $x = a$.

In example 2, the function is discontinuous because condition (3) fails. In example 1, the function is discontinuous because condition (1) fails. The functions in Figs. 129 and 130 are discontinuous because condition (2) fails.

Intuitively, a *function is discontinuous if there is a break of some kind in its graph. A continuous function has an unbroken graph.*

To prove that a given function is continuous is often difficult. But it can be proved that the functions studied in elementary mathematics and its applications are all continuous except possibly for isolated values of x. *Thus polynomials are continuous for all values of x; and rational functions are continuous for all values of x except those for which the denominator becomes zero.* Assuming that these statements are correct we may use part (3) of our definition to find limits of continuous functions; for part (3) of the definition says that the limit of $f(x)$ as x approaches a can be found by substituting a for x in $f(x)$ as long as $f(x)$ is known to be continuous.

Example 3. $\lim_{x \to 0} (7 + 1000\,x + 10000\,x^2) = 7$. Notice that the magnitude of the coefficients of our polynomial has no effect. We have the table:

x	1	.1	.01	.001	.0001	.00001	...
$f(x)$	11007	207	18	8.01	7.1001	7.010001	...

* The formal definition may be omitted without disturbing the continuity of the chapter.

This table makes it intuitively clear that as $x \to 0, f(x) \to 7$. But since we know that a polynomial is continuous for all values of x, we could obtain this limit by merely substituting $x = 0$ in the polynomial.

EXERCISES

Evaluate the following limits:

1. $\lim_{x \to 2} (x^2 - 3)$. **2.** $\lim_{x \to 3} (3x^3 - 8x^2 + 5)$. **3.** $\lim_{x \to 3} \dfrac{x^2 - 9}{x + 5}$.

4. $\lim_{h \to 0} (2x + h)$. **5.** $\lim_{h \to 0} (3x^2 + 2hx + h^2)$. **6.** $\lim_{h \to 0} (h^2 + 3h)$.

Find the value of the expression $\lim_{h \to 0} \dfrac{f(x+h) - f(x)}{h}$ *where:*

7. $f(x) = x^2$. **8.** $f(x) = x^3$. **9.** $f(x) = 4x^2 + 3x - 1$.

10. $f(x) = 5x^3$. **11.** $f(x) = 1/x$. **12.** $f(x) = 3 - 2x - 3x^2$.

13. $f(x) = 1/x^2$.

103. Infinity; how certain functions vary. Consider the function $y = 1/x$ (Fig. 133). It is not defined at $x = 0$. We see that if x takes successively the sequence of values .1, .01, .001, .0001, $\cdots$, $1/10^n$, $\cdots$ then $f(x)$ takes the values 10, 100, 1000, 10000, $\cdots$, 10^n, $\cdots$. Similarly, if x takes the values $-.1$, $-.01$, $-.001$, $\cdots$, $-1/10^n$, $\cdots$ then $f(x)$ takes the values -10, -100, -1000, $\cdots$, -10^n, $\cdots$. Similarly it is easy to see that if $x_1, x_2, x_3, \cdots, x_n$, is any sequence of positive numbers approaching zero as a limit then the terms of the sequence $f(x_1), f(x_2), f(x_3), \cdots, f(x_n), \cdots$ become ultimately larger than any preassigned positive number, no matter how large; and if $x_1, x_2, x_3, \cdots, x_n, \cdots$ is any sequence of negative numbers approaching zero as a limit, then the terms of the sequence $f(x_1), f(x_1), f(x_3), \cdots, f(x_n), \cdots$ become ultimately smaller than any preassigned negative number. In either case the absolute value $|f(x_n)|$ becomes greater than any preassigned positive number. This suggests the following definition.

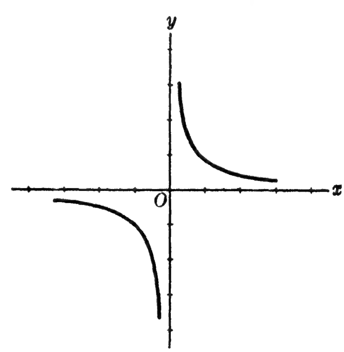


FIG. 133

* DEFINITION. *If for every sequence* $x_1, x_2, \cdots, x_n, \cdots \to a$ *(all* $x_n \neq a$*) and any preassigned number* H *(no matter how large), there exists a corresponding term* x_N *of the sequence such that for all succeeding terms* $x_m (m > N)$ *we have*

(1) $$|f(x_m)| > H$$

then we say that the ***absolute value of f(x) increases indefinitely as x approaches a.*** *In symbols we write* $|f(x)| \uparrow$ *as* $x \to a$.

Intuitively this means that the values of the function $f(x)$ get indefinitely large in absolute value as x gets near to a.

This is often written in the form $\lim\limits_{x \to a} f(x) = \infty$ (read, the limit of $f(x)$ as x approaches a is infinity) or in the form $f(x) \to \infty$ as $x \to a$ (read, $f(x)$ becomes infinite as x approaches a). These symbols sometimes mislead the unwary student into thinking that there is a peculiar number, infinity, which is the limit of $f(x)$ as $x \to a$, or which is approached by $f(x)$ as x approaches a. Therefore we prefer to write $\left|\frac{1}{x}\right| \uparrow$ as $x \to 0$ instead of the usual $\lim\limits_{x \to 0} \frac{1}{x} = \infty$. It would be even more misleading to write $\frac{1}{0} = \infty$, although some books do write it.† Of course, they use this notation to mean that $\left|\frac{1}{x}\right| \uparrow$ as $x \to 0$; they do not mean that ∞ is a number which you obtain by dividing 1 by 0. As we have seen $1/0$ is a meaningless symbol; and furthermore the symbol ∞ is not a number at all. "Infinity," in this sense, is merely a way of describing the manner in which certain functions vary.

Remark. It is sometimes desirable to distinguish between the manner in which the function $1/x$ varies as $x \to 0$ from the right and the behavior of the function as $x \to 0$ from the left. We

* The formal definition may be omitted without disturbing the continuity of the chapter.

† Historically, the statement $1/0 = \infty$ was actually taken literally. But we are now more familiar with the logical foundations of numbers than were many pioneers in the field.

write $\frac{1}{x} \uparrow$ as $x \to 0$ from the right, and $\frac{1}{x} \downarrow$ as $x \to 0$ from the left (read $\frac{1}{x}$ increases indefinitely as x approaches 0 from the right and $\frac{1}{x}$ decreases indefinitely as x approaches 0 from the left). This is often written in the form $\frac{1}{x} \to +\infty$ as $x \to 0$ from the right, and $\frac{1}{x} \to -\infty$ as $x \to 0$ from the left.

EXERCISES

(*a*) *What can be said about the way in which the following functions vary as x approaches the indicated value?*

(*b*) *Draw a graph of each function for values of x near the value in question.*

1. $\dfrac{1}{x-2}$ as $x \to 2$.

2. $\dfrac{17}{2-x}$ as $x \to 2$.

3. $\dfrac{x^2+5x}{x-2}$ as $x \to 2$.

4. $\dfrac{2x^2-3x}{x^2-4}$ as $x \to 2$.

5. $1/x^2$ as $x \to 0$.

6. $-1/x^2$ as $x \to 0$.

104. Tangents. In high school, the tangent to a circle C at a given point P on C was defined as a line which meets the circle

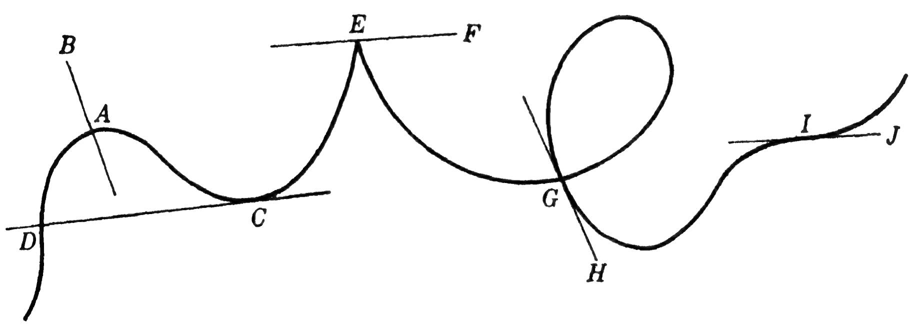


FIG. 134

only at P. This definition is clearly unsuitable for other curves. For (Fig. 134), AB meets the curve only at A but is not what we would like to call a tangent; while CD is what we would like to call a tangent at C but meets the curve again at D. It will not

do to say that a tangent at P is a line which does not "cross" the curve at P for GH and IJ do cross the curve and are what we would like to call tangents, while EF does not cross the curve but is not what we would like to call a tangent. If one thinks of a curve as being generated by a moving particle, then we would like a tangent at a given point on the curve to be a straight line with the same "direction" as that in which the particle is moving at that point. A definition of tangent corresponding to this intuitive idea, will involve the idea of limits. Let P be a point on our curve at which we would like to define a tangent (Fig. 135). Let P_1 be a nearby point on the curve. Then PP_1 is a secant or chord.

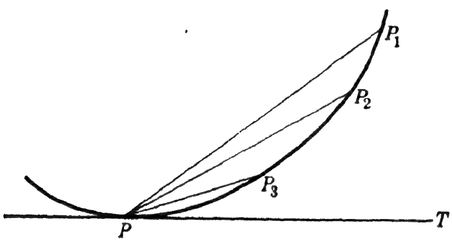


FIG. 135

Now let $P_1, P_2, P_3, \cdots, P_n, \cdots$ be a sequence of points on the curve, all different from P, but approaching P as a limit.* Then we can conceive of the tangent PT at P as the "limiting position" of the sequence of secants $PP_1, PP_2, PP_3, \cdots, PP_n, \cdots$. Actually since "limiting position" has not been defined precisely, it will be more convenient to define the slope of the tangent at P. This is sufficient † because if we know the coordinates of P and the slope of PT we can write the equation of PT (see section 79). Thus we shall define *the slope of the tangent at P as the limit of the slopes of the secants PP_n as $P_n \to P$ (if this limit exists).* Note that according to this definition, the "tangent" to a straight line at any point of the line is the line itself. This definition will be studied more closely in the next section.

105. The derivative of a function. Consider the function $y = f(x)$ at $x = x_1$. Let $y_1 = f(x_1)$. Take a nearby point (x_2,y_2) on the curve (Fig. 136); $y_2 = f(x_2)$. The slope of the secant P_1P_2 is $\dfrac{y_2 - y_1}{x_2 - x_1}$. Let $x_2 - x_1$ be called Δx, read "delta x," or the

* This means that the distances $PP_1, PP_2, PP_3, \cdots, PP_n, \cdots$ approach zero as a limit.

† Except for vertical tangents which will not be discussed here.

difference in x or the **change in x** or the **increment in x** or the **increase in x.** Similarly, let $y_2 - y_1$ be called Δy, or the **change in y.** (The symbol Δx is not to be understood as a number Δ multiplied by a number x; Δx stands for a single quantity which happens to have a first and second name, just as you have.) Then the slope of the secant is $\dfrac{\Delta y}{\Delta x}$. As P_2 approaches P_1, x_2 approaches x_1, and $\Delta x \to 0$; similarly $y_2 \to y_1$ and $\Delta y \to 0$. But it may happen that $\dfrac{\Delta y}{\Delta x}$ approaches a limit. (See example 1, section 102.) If this limit exists, it is called the **slope of the tangent at (x_1,y_1).**

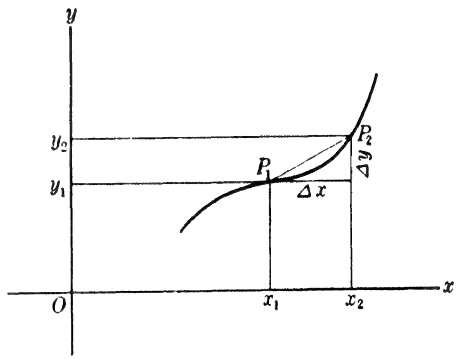


FIG. 136

Example. Consider the curve $y = x^2$. See Fig. 137. Let us find the slope of the tangent at $x_1 = 3$, $y_1 = 9$. Take a nearby value of x, say $x_2 = 3 + \Delta x$. Then for this value of x, $y_2 = (3 + \Delta x)^2 = 9 + 6 \cdot \Delta x + (\Delta x)^2$. Now Δy means, by definition, the difference between this value y_2 and the value $y_1 = 9$. Thus

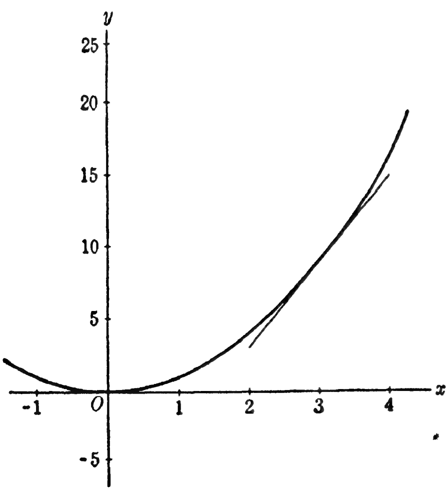


FIG. 137

$$(1) \qquad \Delta y = 6 \cdot \Delta x + (\Delta x)^2.$$

Dividing both sides of (1) by Δx, we find that the slope of the secant is

$$\frac{\Delta y}{\Delta x} = 6 + \Delta x.$$

Therefore as $\Delta x \to 0$ we find that the slope of the tangent at (3,9) is

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 6.$$

It might be worth while to make a table for this example which will make the idea clear:

x_1	$x_2 = x_1 + \Delta x$	$\Delta x = x_2 - x_1$	y_1	$y_2 = y_1 + \Delta y = f(x_1 + \Delta x)$	Δy	$\frac{\Delta y}{\Delta x}$
3	5	2	9	25	16	8
3	4	1	9	16	7	7
3	3.1	.1	9	9.61	.61	6.1
3	3.01	.01	9	9.0601	.0601	6.01
3	3.001	.001	9	9.006001	.006001	6.001

From this table it is intuitively clear that as Δx gets closer and closer to zero, Δy also approaches zero, but $\frac{\Delta y}{\Delta x}$ approaches 6.

It is usually possible to evaluate the slope of the tangent for all values of x at once. Consider the example $y = x^2$ again. Take any point (x,y) on the curve. Take a nearby value of x; call it $x + \Delta x$. The ordinate of the point on the curve corresponding to this abscissa is $(x + \Delta x)^2$; call it $y + \Delta y$. Thus

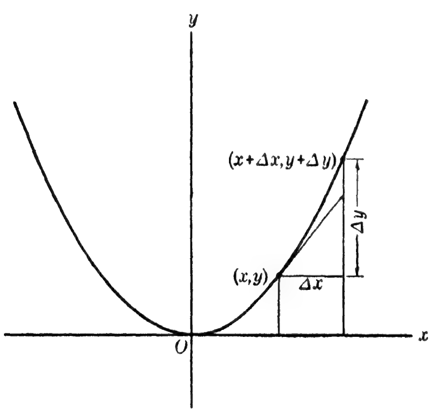


FIG. 138

(2) $$y + \Delta y = x^2 + 2x \cdot \Delta x + (\Delta x)^2.$$

But

(3) $$y = x^2.$$

Hence, subtracting (3) from (2), we have

(4) $$\Delta y = 2\,x \cdot \Delta x + (\Delta x)^2.$$

Therefore, dividing both sides of (4) by Δx, we have $\frac{\Delta y}{\Delta x} = 2\,x + \Delta x$, which is the slope of the secant joining (x,y) and $(x + \Delta x, y + \Delta y)$. Thus the slope of the tangent at (x,y) is $\lim\limits_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 2\,x$. This is true for any value of x. For example,

for $x = 3$, the slope of the tangent is $2 \cdot 3 = 6$, as we saw before. Notice that the slope of the tangent to the curve $y = x^2$ at the point whose abscissa is x is $2x$. This is a new function of x, called the *derived function or derivative of the function* x^2. In general, we make the following definition.

DEFINITION. *Consider the function* $y = f(x)$. *Let* $y + \Delta y = f(x + \Delta x)$ *for any increment* Δx; *then* $\Delta y = f(x + \Delta x) - f(x)$. *Now,*

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$

is a new function of x *(provided this limit exists) called the* ***derivative*** *or* ***derived function of*** $f(x)$. *The derivative of the function* $y = f(x)$ *is denoted by* $\frac{dy}{dx}$, *read "the derivative of* y *with respect to* x," *or* $\frac{d}{dx}f(x)$, *read "the derivative of* $f(x)$ *with respect to* x," *or* $\frac{df}{dx}$, *read "the derivative of* f *with respect to* x," *or* $f'(x)$, *read "*f *prime of* x," *or "the derivative of* $f(x)$," *or simply* y', *read "*y *prime" or "the derivative of* y." *Briefly,*

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}.$$

Remark 1. We have assumed that the limit $\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$ exists. Now

$$\frac{\Delta y}{\Delta x} = \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

is, for a fixed x, a function of the independent variable Δx. We have seen that the limit of a function may not exist. But it can be proved that for the functions studied in elementary mathematics and its applications this limit always does exist, except possibly for isolated values of x.

Remark 2. The symbol $\frac{dy}{dx}$ should not be thought of as the quotient of a quantity dy divided by a quantity dx. In particular, we cannot regard dx as meaning $\lim_{\Delta x \to 0} (\Delta x)$ and dy as meaning $\lim_{\Delta x \to 0} (\Delta y)$, since each of these limits is clearly zero and $\frac{dy}{dx}$ would then become $0/0$ which is undefined. Misinterpretations

like this caused considerable trouble during the 17th and 18th centuries, since mathematicians, inspired by the tremendous success of the calculus in solving problems in both pure and applied mathematics, devoted their energies to developing the powerful methods of the subject and did not straighten out its logical foundations until the 19th century. Bishop Berkeley, in the 18th century, scoffed at derivatives as quotients of "the ghosts of departed quantities." This is reminiscent of the grinning Cheshire cat, in Lewis Carroll's *Alice in Wonderland*, who disappeared gradually, beginning with the end of his tail and working forward until nothing was left but the grin. However, it is not surprising, when viewed correctly, that two functions may each approach zero while their quotient approaches a value different from zero. For example, $\lim\limits_{h \to 0} (6h) = 0$ and $\lim\limits_{h \to 0} (3h) = 0$ but $\lim\limits_{h \to 0} \left(\frac{6\,h}{3\,h}\right) = \frac{6}{3} = 2$. Berkeley wrote* that a person who can digest some of the concepts of the calculus "need not, methinks, be squeamish about any point in divinity." Berkeley's criticisms were not without foundation and helped to stimulate mathematicians to examine critically the logical foundations of the subject. For example, Berkeley asked "whether certain maxims do not pass current among analysts which are shocking to good sense? And whether the common assumption, that a finite quantity divided by nothing is infinite, be not of this number?" Thus Berkeley justly balked at the expression "$1/0 = \infty$" which was discussed in section 103. It was not until the 19th century that mathematicians began to avoid writing this misleading expression.

The process of finding the derivative of a given function is called **differentiation.** The study of derivatives is called the **differential calculus.** As we have already pointed out, *the value of the derivative $f'(x)$ at $x = x_1$ is the slope of the tangent to the curve $y = f(x)$ at (x_1,y_1) where $y_1 = f(x_1)$.*

EXERCISES

In each of the following exercises, (a) find the derivative of the given function; (b) find the slope of the tangent to the curve at the point where $x = 2$; (c) write the equation

* *The Analyst*, London, 1734.

of the tangent to the curve at the point where $x = 2$; (d) plot the curve and the tangent line found in part (c):

1. $y = 3x^2$. **2.** $y = 2x^2$. **3.** $y = \frac{1}{2}x^2$.

4. $y = 3x^2 - 5x + 6$. **5.** $y = 2x^2 + 5$. **6.** $y = x^3$.

7. $y = 2x^2 - 3x + 5$. **8.** $y = 1/x$. **9.** $y = 1/x^2$.

10. $y = 3x$. **11.** $y = 2x + 3$. **12.** $y = x - 2$.

13. $y = x$.

14. Find the derivative of $y = mx + p$. Interpret graphically.

15. Find the derivative of $y = cx^2 + kx + a$.

106. Instantaneous rate of change. We have seen in section 93 that if $y = f(x)$ then the average rate of change of y with respect to x for the interval from x_1 to x_2 is the slope of the secant joining the points (x_1,y_1) and (x_2,y_2) of the curve $y = f(x)$. This average rate of change is $\frac{\Delta y}{\Delta x}$. If y varied with x at a uniform rate then we might say that the rate of change at any instant is the same as the average rate for any interval, since all these average rates would be the same. But if y varies with x at a changing rate, what shall we mean by the rate of change of y with respect to x at a given value of x? For example, a falling body does not fall at a uniform speed but falls faster and faster as it falls. If s is the distance through which the body falls in t seconds then s is a function of t and the "speed" with which the body falls is the rate of change of s with respect to t. What shall we mean by the speed of a falling body at a given instant (value of t)? Clearly the average speed for a very small interval of time beginning with the given instant will be very close to what we want to mean by the instantaneous speed at that instant. This suggests the following definition.

DEFINITION. *If $y = f(x)$, the **instantaneous rate of change of y with respect to x, at x** $= x_1$, is $\lim\limits_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$, or $f'(x_1)$ where $f'(x)$ is the derivative $\frac{dy}{dx}$ of y with respect to x.*

Hence, an instantaneous rate corresponds graphically to the slope of a tangent, just as an average rate corresponds graphically to the slope of a secant.

Example. Consider the falling body problem. Here $s = 16\,t^2$ where s is the distance, measured in feet, through which the body falls in t seconds. What is the instantaneous speed of the falling body at the end of the first second? We have following table:

t	$t + \Delta t$	Δt	s	$s + \Delta s = 16(t + \Delta t)^2$	Δs	$\frac{\Delta s}{\Delta t}$ = *average speed for the interval from t to $t + \Delta t$*
1	2	1	16	64	48	48
1	1.5	.5	16	36	20	40
1	1.1	.1	16	19.36	3.36	33.6
1	1.01	.01	16	16.3216	.3216	32.16
1	1.001	.001	16	16.032016	.032016	32.016

This table indicates that the average speed approaches 32 ft. per sec. as a limit as $\Delta t \to 0$. This can be proved as follows. We have

$$s + \Delta s = 16(t + \Delta t)^2$$

or

$$s + \Delta s = 16\,t^2 + 32 \cdot t \cdot \Delta t + 16\,(\Delta t)^2.$$

Then

$$\Delta s = 32\,t \cdot \Delta t + 16\,(\Delta t)^2$$

and the average speed for the interval Δt is

$$\frac{\Delta s}{\Delta t} = 32\,t + 16\,\Delta t.$$

Hence the instantaneous speed at the end of t seconds is

$$\lim_{\Delta t \to 0} \frac{\Delta s}{\Delta t} = 32\,t$$

or

$$\frac{ds}{dt} = 32\,t \text{ ft. per sec.}$$

At $t = 1$, this speed is 32 ft. per sec. as suggested by the table.

The idea of an instantaneous rate of change is very useful in physics. Thus if a rectilinear motion (that is, motion in a straight line) is specified by giving the distance s from the starting point as a function of the time t, then $\frac{ds}{dt}$ is the **velocity** v. The

rate of change of the velocity with respect to time, or $\frac{dv}{dt}$ is called the **acceleration.** In the above example, $v = 32\,t$. Hence, the acceleration of a falling body is $\frac{dv}{dt}$ which may be calculated as follows:

$$v + \Delta v = 32(t + \Delta t) = 32\,t + 32 \cdot \Delta t$$
$$\Delta v = 32\,\Delta t$$
$$\frac{\Delta v}{\Delta t} = 32$$
$$\frac{dv}{dt} = \lim_{\Delta t \to 0} \frac{\Delta v}{\Delta t} = 32 \text{ ft. per sec. per sec.}$$

This is a constant and does not depend on the value of t; hence the acceleration of a falling body is 32 ft. per sec. per sec. at any instant.

The problems of tangents and of instantaneous rates of change like velocity and acceleration were actually what led to the idea of the derivative of a function. Here again we see how the abstract idea of the derivative of a function may be applied to or interpreted in different concrete situations such as tangents in geometry or velocity and acceleration in physics. Newton was led to the invention of the differential calculus, or "method of fluxions" as he called it, because he was concerned with ideas like velocity and acceleration. Much of the importance and power of this calculus is due to the fact that it enables us to depart from static considerations and to study the dynamical problems of a changing, moving world.

EXERCISES

1. If a projectile is thrown straight up from the ground with an initial velocity of 96 ft. per sec., its height h above the ground is given by $h = 96\,t - 16\,t^2$ where h is measured in feet and t in seconds. (*a*) What is the velocity at $t = 1, 2, 3, 4$? (*b*) What is the acceleration at these times? (*c*) Interpret the plus and minus signs in your answers. (*d*) Draw a graph of this function from $t = 0$ to $t = 6$, plotting t on the horizontal axis and h on the vertical axis. Describe the motion of the projectile. (*e*) Is the graph of the motion a picture of the *path* of the projectile?

2. The distance s feet a ball rolled in a straight line down an inclined plane at the end of t seconds is given by $s = 6\,t^2$. (*a*) What is its average speed for the

first two seconds? (*b*) What is its instantaneous speed at $t = 2$? (*c*) Find the acceleration at any instant.

3. A ball starts rolling up an inclined plane, moving in a straight line. Its distance from the starting point is given by $s = 100\,t - 10\,t^2$ where s is measured in feet and t in seconds. (*a*) Find expressions for the velocity and acceleration at any instant. (*b*) What is the velocity at $t = 5$? $t = 10$? $t = 15$?

4. The volume of a spherical balloon is given by $v = \frac{4}{3}\pi r^3$ where r is the radius and $\pi = 3.14 \cdots$. Suppose the balloon is being blown up. Find the rate of change of the volume with respect to the radius.

5. Find the rate of change of the area of a circle with respect to the radius, recalling that $A = \pi r^2$.

107. Differentiation of polynomials. We have been differentiating each function considered by appealing directly to the definition of derivative. In this section we shall learn how to differentiate polynomials and certain other functions quickly.

Consider the function

$$(1) \qquad y = x^n$$

where n is a natural number greater than one. Then $y + \Delta y = (x + \Delta x)^n$. By direct multiplication we find * that

$$\begin{aligned}
(x + \Delta x)^2 &= x^2 + 2\,x \cdot \Delta x + (\Delta x)^2 \\
(x + \Delta x)^3 &= x^3 + 3\,x^2 \cdot \Delta x + 3\,x(\Delta x)^2 + (\Delta x)^3 \\
(x + \Delta x)^4 &= x^4 + 4\,x^3\Delta x + 6\,x^2(\Delta x)^2 + 4\,x(\Delta x)^3 + (\Delta x)^4 \\
(x + \Delta x)^5 &= x^5 + 5\,x^4\Delta x + 10\,x^3(\Delta x)^2 + 10\,x^2(\Delta x)^3 + \\
&\qquad 5\,x(\Delta x)^4 + (\Delta x)^5
\end{aligned}$$

and so on. These experiments suggest that for any natural number $n > 1$,

$$(2) \quad y + \Delta y = (x + \Delta x)^n = x^n + nx^{n-1}\Delta x + \text{terms involving } \Delta x \text{ with an exponent of at least 2.}$$

This could be proved rigorously, but we shall take it for granted † here. Then subtracting (1) from (2), we obtain

$$(3) \qquad \Delta y = nx^{n-1}\Delta x + \text{terms involving } \Delta x \text{ with an exponent of at least 2.}$$

Hence, dividing both sides of (3) by Δx, we have

$$\frac{\Delta y}{\Delta x} = nx^{n-1} + \text{terms involving } \Delta x \text{ with an exponent of at least 1.}$$

* The student should verify the results of these multiplications.

† See Appendix, section 173.

Thus

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = nx^{n-1}.$$

This proves the following theorem.

THEOREM 1. *If $y = x^n$, where n is any natural number greater than one, then* $\frac{dy}{dx} = nx^{n-1}$*; that is, for $n > 1$,*

(4) $$\frac{d}{dx}(x^n) = nx^{n-1}.$$

We have already seen (see exercise 13, section 105) that the derivative of x^1 with respect to x is 1. Hence (4) holds even for $n = 1$ since it says that the derivative of x^1 with respect to x is $1 \cdot x^0$ which *is* equal to 1. Hence *formula (4) is valid for all natural numbers n.*

THEOREM 2. *Formula (4) is valid when n is any real number whatever, positive, negative or zero.*

We shall not prove this here.

Example 1. The derivative of x^5 with respect to x is $5x^4$ by (4).

Example 2. Find the derivative of $1/x^2$. Now $1/x^2 = x^{-2}$. By (4), the derivative of x^{-2} is $-2x^{-3}$. Hence $\frac{d}{dx}(1/x^2) = -2/x^3$.

THEOREM 3. *The derivative of a constant c with respect to x is zero.*

Proof. If $y = c$ then $y + \Delta y = c$ and hence $\Delta y = 0$. Therefore $\frac{\Delta y}{\Delta x} = 0$ and $\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = 0$.

Graphically, this means that the slope of the straight line $y = c$ is zero (that is, the line is horizontal) since the tangent to any straight line is the line itself.

We shall use the following theorems without proof.

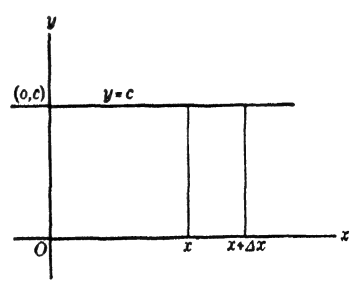


FIG. 139

THEOREM 4. *The derivative of a product of any constant c times any*

function $f(x)$ is equal to c times the derivative of $f(x)$. That is,

$$\frac{d}{dx}[c \cdot f(x)] = c \cdot \frac{d}{dx} f(x). \tag{5}$$

THEOREM 5. *The derivative of the sum of any two functions $f(x)$ and $g(x)$ is equal to the sum of their derivatives. That is,*

$$\frac{d}{dx}[f(x) + g(x)] = \frac{d}{dx} f(x) + \frac{d}{dx} g(x). \tag{6}$$

By using the above theorems, we can differentiate any polynomial and certain other functions as in the following examples.

Example 3. Differentiate the polynomial $6x^3 + 3x^2 + 5$. By Theorem 5,

$$\frac{d}{dx}(6x^3 + 3x^2 + 5) = \frac{d}{dx}(6x^3) + \frac{d}{dx}(3x^2) + \frac{d}{dx}(5).$$

By Theorem 4, $\frac{d}{dx}(6x^3) = 6 \cdot \frac{d}{dx}(x^3)$ and $\frac{d}{dx}(3x^2) = 3 \cdot \frac{d}{dx}(x^2)$.

By Theorem 3, $\frac{d}{dx}(5) = 0$. Hence,

$$\begin{aligned} \frac{d}{dx}(6x^3 + 3x^2 + 5) &= 6 \cdot \frac{d}{dx}(x^3) + 3 \cdot \frac{d}{dx}(x^2) + 0 \\ &= 6 \cdot 3x^2 + 3 \cdot 2x \qquad \text{(by Theorem 1)} \\ &= 18x^2 + 6x. \end{aligned}$$

Example 4. Differentiate $\frac{3}{x^2} + \frac{4}{x^3}$. We have

$$\begin{aligned} \frac{d}{dx}\left(\frac{3}{x^2} + \frac{4}{x^3}\right) &= \frac{d}{dx}\left(3 \cdot \frac{1}{x^2}\right) + \frac{d}{dx}\left(4 \cdot \frac{1}{x^3}\right) && \text{(by Theorem 5)} \\ &= 3 \cdot \frac{d}{dx}\left(\frac{1}{x^2}\right) + 4 \cdot \frac{d}{dx}\left(\frac{1}{x^3}\right) && \text{(by Theorem 4)} \\ &= 3 \cdot \frac{d}{dx}(x^{-2}) + 4 \cdot \frac{d}{dx}(x^{-3}) \\ &= 3(-2)x^{-3} + 4(-3)x^{-4} && \text{(by Theorem 2)} \\ &= -6x^{-3} - 12x^{-4} = -\frac{6}{x^3} - \frac{12}{x^4}. \end{aligned}$$

EXERCISES

Differentiate each of the following functions by means of Theorems 1–5 above:

1. $4x^3 + 6x^2$. **2.** $5x^3 + 2x^2 - 3x + 1$. **3.** $x^2 - 5x$.

4. $2x + 3$. **5.** $5x^4 + 2x^3 - 3x^2 - 4x - 5$.

6. $\dfrac{2}{x^4} + \dfrac{3}{x^2}$. 7. $2x^3 + \dfrac{2}{x^3}$. 8. $3x^2 + 5 + \dfrac{2}{x}$.

9. $2x - 3 + \dfrac{4}{x^2}$.

In each of the following (a) find the slope of the tangent to the curve at the indicated point; (b) write the equation of the tangent to the curve at the indicated point; (c) plot the curve and the tangent found in part (b):

10. $y = x^2 - 3$ at the point for which $x = 2$.
11. $y = x^3$ at the point for which $x = 1$.
12. $y = x^3$ at the point for which $x = -1$.
13. $y = x^2 - 2x + 3$ at the point for which $x = 1$.
14. $y = 9 - x^2$ at the point for which $x = 3$.
15. $y = 9 - x^2$ at the point for which $x = -3$.
16. $y = 12/x$ at the point for which $x = 2$.
17. $y = 12/x$ at the point for which $x = -2$.
18. $y = 1/x^2$ at the point for which $x = 2$.
19. $y = 1/x^2$ at the point for which $x = -2$.

20. For what value of x is the tangent to the curve $y = x^2 - 4x + 5$ horizontal? Plot.

21. For what values of x are the tangents to the curve $y = \frac{1}{3}x^3 - \frac{1}{2}x^2 - 6x + 2$ horizontal? Plot.

22. (a) Differentiate x^2. (b) Differentiate x^3. (c) Differentiate x^5. (d) Is it true that the derivative of the product of two functions is equal to the product of their derivatives?

23. (a) Differentiate x^6. (b) Differentiate x^2. (c) Differentiate x^4. (d) Is it true that the derivative of the quotient of two functions is equal to the quotient of their derivatives?

24. A falling body traverses s ft. in t seconds where $s = 16t^2$. (a) Find the velocity and acceleration at any instant. (b) At $t = 2$. (c) At $t = 5$.

25. (a) How fast does the volume of a cube change with respect to its edge x? (b) What is the rate of change of its surface area with respect to its edge x?

108. Maxima and minima. Consider the curve $y = f(x)$ and let $f'(x)$ be the derivative of y with respect to x. Recall that $f'(x)$ is the slope of the tangent at the point whose abscissa is x. Let us consider how y changes as x goes from left to right (increases). If $f'(x)$ is positive, clearly y is rising at that point; if $f'(x)$ is negative, y is falling; if $f'(x) = 0$, the tangent is horizontal (Fig. 140). A value of x for which the tangent is horizontal is called a **critical value.** A point of the curve is

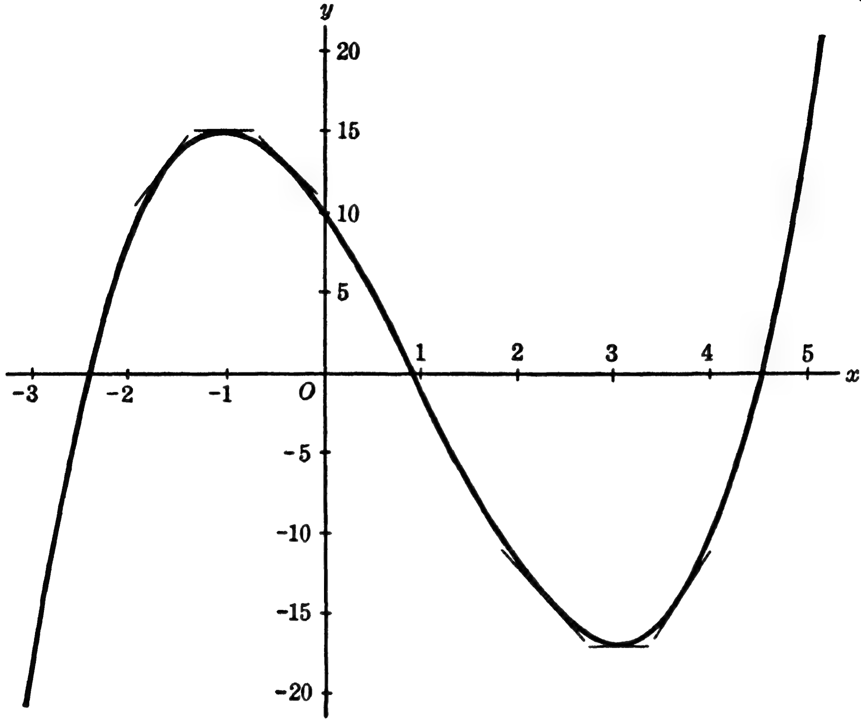


FIG. 140

called a **maximum** if it is higher than any nearby point; **minimum** if it is lower than any nearby point. If (a,b) is a maximum point, the function $f(x)$ is said to have a **maximum value at $x = a$.** If (a,b) is a minimum point the function $f(x)$ is said to have a **minimum value at $x = a$.** In any "smooth" curve, it is intuitively clear that at any maximum or minimum point, the curve has a horizontal tangent. (It can be proved that this is so for all the functions studied in elementary mathematics. This is not so for curves with "corners" like that in the accompanying Fig. 141.)

FIG. 141

But the converse is not true; there are points (Fig. 142) which have a horizontal tangent and which are neither maxima nor minima. It is easy to see that a function $f(x)$ has a maximum at $x = a$ if $f'(x)$ is positive for a value of x slightly less than a and negative for a value of x slightly greater than a. Similarly, $f(x)$

has a minimum at $x = a$, if $f'(x)$ is negative for a value of x slightly to the left of a and positive for a value of x slightly to the right of a. Thus we have the following working rule for finding the maxima and minima of a function:

1. Find the derivative $f'(x)$ of the function.

2. Find the critical values of x; that is, the values of x for which $f'(x) = 0$.

3. For each critical value $x = a$, evaluate $f'(x)$ for values of x slightly to the left and right of a. If $f'(x)$ changes from positive to negative as we go from left to right, $f(x)$ has a maximum at $x = a$. If $f'(x)$ changes from negative to positive as we go from left to right, $f(x)$ has a minimum at $x = a$. If $f'(x)$ does not change sign as we go from left to right, $f(x)$ has neither a maximum nor a minimum at $x = a$.

Example 1. Test the curve $y = x^3 - 3x^2 - 9x + 10$ for maxima and minima. Differentiating, we obtain

$$\frac{dy}{dx} = f'(x) = 3x^2 - 6x - 9.$$

The critical values are obtained by solving the equation $3x^2 - 6x - 9 = 0$, or $x^2 - 2x - 3 = 0$, whose roots are $x = 3$ and $x = -1$. We may conveniently tabulate our work as follows:

x	2	3	4	-2	-1	0
$\frac{dy}{dx} = f'(x)$	neg.	0	pos.	pos.	0	neg.
$y = f(x)$		-17			15	

Hence at $x = 3$, the function y attains a minimum value of -17 and at $x = -1$ a maximum value of 15. See Fig. 140.

Example 2. Test the curve $y = x^3 - 3x^2 + 3x + 1$ for maxima and minima. Differentiating, we obtain $f'(x) = 3x^2 - 6x + 3$. Thus the only critical value is the root of $3x^2 - 6x + 3 = 0$, or $x^2 - 2x + 1 = 0$, or $x = 1$. We have

x	0	1	2
$\frac{dy}{dx} = f'(x)$	pos.	0	pos.
$y = f(x)$		2	

Hence the critical value $x = 1$ yields neither a maximum nor a minimum. See Fig. 142.

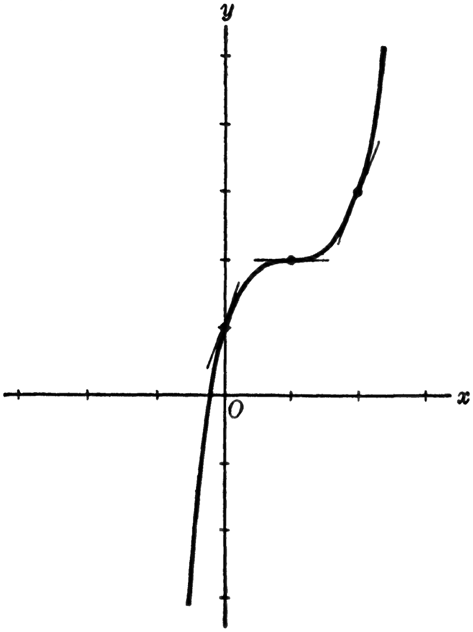


FIG. 142

Remark. Note that a function may attain higher values than it does at a maximum value and lower values than it does at a minimum. For example, the function in example 1 attains the value 620 at $x = 10$; this is greater than the "maximum" value 15. These terms are used relative to sufficiently nearby points of the curve. Any peak in a mountain range is a maximum point even though it is not the highest peak of all. In the usual technical language, we are here finding "relative maxima and minima" rather than "absolute maxima and minima."

EXERCISES

Find the maxima and minima of the following curves and plot:

1. $y = x^2 - 6x + 2.$
2. $y = 6x - x^2 + 2.$
3. $y = 4x - x^2 + 3.$
4. $y = x^2 + 4x + 3.$
5. $y = 2x^3 - 9x^2 - 24x - 12.$
6. $y = \frac{1}{3}x^3 - \frac{1}{2}x^2 - 2x + 5.$
7. $y = x^3 - 6x^2 + 12x - 8.$
8. $y = \frac{1}{3}x^3 + \frac{1}{2}x^2 - 2x + 3.$

109. Applied maxima and minima. Problems involving maxima and minima are obviously among the most important practical considerations that arise in applied mathematics. An engineer, manufacturer, or other practical worker tries to get the best effect with the least effort, or the most efficient result with the least cost. Such problems often amount mathematically to the

investigation of the maxima and minima of some function. From the conditions of the problem, we must first express the quantity in which we are interested as a function of one variable.

Example 1. A rectangular area is to be enclosed by 40 feet of fence. What should the dimensions of the rectangle be if we want the greatest possible area to be enclosed? What is the maximum area?

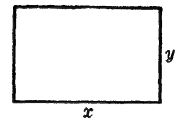


FIG. 143

Clearly, we might have many rectangles, 2 by 18, 3 by 17, 4 by 16, 4.5 by 15.5, etc. with different areas, all having the required perimeter, 40 ft. Let x and y be the length and width of any rectangle. Then $2x + 2y = 40$ by hypothesis. Hence $y = 20 - x$. Now the area A of the rectangle is given by $A = xy$, or, $A = x(20 - x)$, or $A = 20x - x^2$. It is this particular function of x of which we want the maximum value. Hence we find $\frac{dA}{dx} = 20 - 2x$. This derivative is zero only at $x = 10$. To the left of this critical value, say at $x = 9$, $\frac{dA}{dx}$ is positive and to the right, say at $x = 11$, it is negative. Hence $x = 10$ gives a maximum. Thus the rectangle 10 by 10 has the greatest possible area, namely 100 sq. ft. It will be instructive to plot the graph of this function $A = 20x - x^2$ (Fig. 144) and to observe the graphical significance of the facts just discussed.

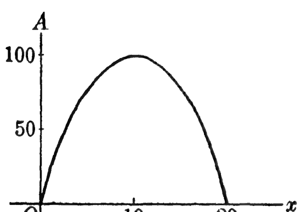


FIG. 144

Example 2. An open box is to be made from a square piece of cardboard 12 inches on each side by cutting out equal squares from the corners and folding up the sides. How long should the edge of the cut-out square be in order to obtain a box of maximum volume? What is the maximum volume? See Fig. 145.

Let x be the length of the edge of the cut-out square. Then the volume of the box is given by $V = x(12 - 2x)^2 = 144x - 48x^2 + 4x^3$. This is the function of which we want the maximum value. We find $\frac{dV}{dx} = 144 - 96x + 12x^2$. The critical

values of x are the roots of the equation

$$144 - 96\,x + 12\,x^2 = 0,$$

or

$$12 - 8\,x + x^2 = 0,$$

or

$$(x - 6)(x - 2) = 0,$$

or

$$x = 6,\ x = 2.$$

The critical value $x = 6$ need not be considered since, for this value, there would be no box at all. The value $x = 2$ gives a maximum since for $x = 1$, $\frac{dV}{dx}$ is positive and for $x = 3$ it is negative. The box 2 by 8 by 8 has the maximum volume 128 cu. in.

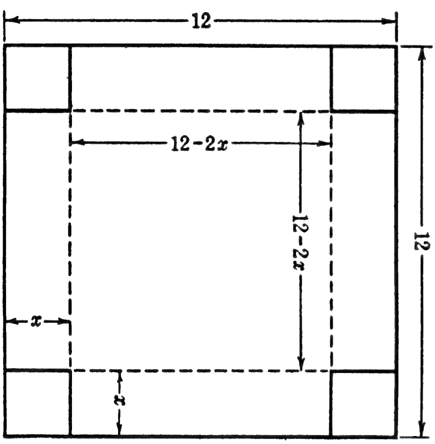


FIG. 145

Example 3. A projectile is thrown straight up from a height of 5 ft., at an initial velocity of 128 ft. per sec. Its height above the ground after t seconds is given by $h = 5 + 128\,t - 16\,t^2$, where h is measured in ft. What is its maximum height and after how many seconds is it reached?

Differentiating, we obtain

$$\frac{dh}{dt} = 128 - 32\,t.$$

The only critical value is $t = 4$, which gives a maximum height of 261 ft.

EXERCISES

1. A rectangular area is to be enclosed by 100 ft. of fence. Find the dimensions of the rectangle which will have the greatest area. What is the maximum area?

2. An open box is to be made of a square piece of tin 18 inches on each side by cutting out equal squares from the corners and folding up the sides. How long should the side of the cut-out square be in order to get a box with the greatest possible volume? What is the maximum volume?

3. A projectile is thrown straight up from the ground with an initial velocity of 96 ft. per sec. Its height after t seconds is given by $h = 96\,t - 16\,t^2$, where h is measured in feet. After how many seconds does the projectile reach its maximum height? What is the maximum height?

4. Find the dimensions of the rectangle with the smallest perimeter which has an area of 169 sq. ft. (Hint: make use of negative exponents.)

5. A closed box having a volume of 64 cu. ft. is to have a square base. If the total area of the top, bottom, and 4 sides is to be minimum, find the dimensions of the box.

6. A closed box having a volume of 64 cu. ft. is to have a square base. If the material for the top and bottom costs 16 cents per sq. ft. and the material for the sides costs 2 cents per sq. ft., what should be the dimensions of the box in order to make the cost a minimum?

7. A projectile is thrown straight up from a height of 6 ft. with an initial velocity of 192 ft. per sec. Its height after t seconds is given by $h = 6 + 192\,t - 16\,t^2$ where h is measured in feet. After how many seconds does the projectile reach its maximum height? What is the maximum height?

8. It is desired to fence off a rectangular field 3200 sq. yds. in area along the straight bank of a river. No fence is needed along the river bank. What dimensions should the field have if the amount of fencing used is to be a minimum?

9. A printed page is to contain 432 sq. cm. of actual printed matter. There is to be a margin 4 cm. wide along the sides and 3 cm. wide along the top and bottom. What should be the dimensions of the page if the amount of paper used is to be a minimum?

10. An open box with a square base is to be made from 400 sq. in. of material. Find the dimensions of the box with maximum volume. What is the maximum volume?

11. An open box is to have a rectangular base twice as long as it is wide. If the box is to be made of 600 sq. in. of material, find the dimensions which will yield the maximum volume. What is the maximum volume?

12. If 400 people will attend a moving picture theater when the admission price is 30 cents and if the attendance decreases by 40 for each 10 cents added to the price, then what price of admission will yield the greatest gross receipts?

13. A telephone company can get 1000 subscribers at a monthly rate of \$5.00 each. It will get 100 more subscribers for each 10 cent decrease in the rate. What rate will yield the maximum gross monthly income and what will this income be?

14. For a box with square ends to be sent by parcel post, the sum of its length and girth (perimeter of a cross-section) must not exceed 84 inches. Find the dimensions of the box with maximum volume that can be sent. What is the maximum volume?

110. The numbers *e* and π. The study of derivatives, called the differential calculus, is based on the idea of limits as we have seen. This is not the only way that limits enter into mathematics. In fact, every irrational number may be considered as the limit of a sequence of rational numbers. Thus $\sqrt{2}$ is the

limit of the sequence 1, 1.4, 1.41, 1.414, $\cdots$ and π is the limit of the sequence 3, 3.1, 3.14, 3.141, 3.1415, 3.14159, $\cdots$.

An interesting and important irrational number called e arises from the compound interest law. If 1 dollar is invested at 100% interest compounded annually for one year, the amount at the end of the year is clearly 2 dollars. If the interest is compounded semi-annually, the amount is given by $A = 1\left(1 + \frac{1}{2}\right)^2 = 2.25$ dollars. If it were compounded quarterly, the amount would be $A = \left(1 + \frac{1}{4}\right)^4 = 2.441$ dollars approximately. Clearly the more often it is compounded, the greater the amount will be at the end of the year. How great can the amount become if we allow the interest to be compounded oftener and oftener, as every day or every minute or every tenth of a second? Intuition would probably lead us to expect a fortune, but the answer is, surprisingly enough, that the amount cannot surpass 2.72 dollars no matter how often the interest is compounded. This is so because $(1 + 1/n)^n$ is the amount obtained by compounding the interest n times during the year and it can be proved that

$$\lim_{n\uparrow} (1 + 1/n)^n = 2.71828 \cdots. \tag{1}$$

The number defined by (1) is called e and is used as the base of the so-called "natural" logarithms. We cannot prove (1) here but it can be made plausible by the following table:

n	1	2	4	10	100	1000	$\cdots$
$(1 + 1/n)^n$	2	2.25	2.4414 $\cdots$	2.5937 $\cdots$	2.7048 $\cdots$	2.7169 $\cdots$	$\cdots$

The irrational numbers e and π are perhaps the strangest members of number society. They enter into advanced mathematics in many curious and important ways.

The number $\pi = 3.14159265 \cdots$, approximately, is familiar to you but you have probably never learned how it may be calculated. Notice that $\pi \neq 22/7$; the number $22/7 = 3.142 \cdots$ and therefore gives the value of π correctly only as far as the hundredths place. It is interesting to note that, as far as we know, Euclid never calculated π. Euclid did prove the theorem that if A is the area of a circle of radius r and A' is the area of a

circle of radius r' then $A/r^2 = A'/r'^2$. That is, the area of a circle divided by the square of the radius gives the same numerical result no matter how large or small a circle we take. This number A/r^2 may be called π for short; hence we have the familiar formula $A = \pi r^2$ for the area of any circle. Euclid, a pure mathematician, interested chiefly in the logical proof of his theorems, never estimated the value of π. Various crude estimates had been made before but no method, so far as we know, was developed for obtaining this value, possibly because there was no great practical need for it. For example, the Old Testament's estimate * of 3 for the value of π may have been good enough for the roughly circular ox-cartwheels of that era.† It would be hopelessly inadequate for the wheels of an automobile in which to ride comfortably on smooth roads. It remained for Archimedes, a brilliant mathematician and physicist who lived in the century following Euclid, to develop such a method. It can be done in the following way. Take a circle of radius one; then π is exactly the area of the circle. Inscribe a square in the circle and circumscribe a square about it. Then it is easy to see (Fig. 146)

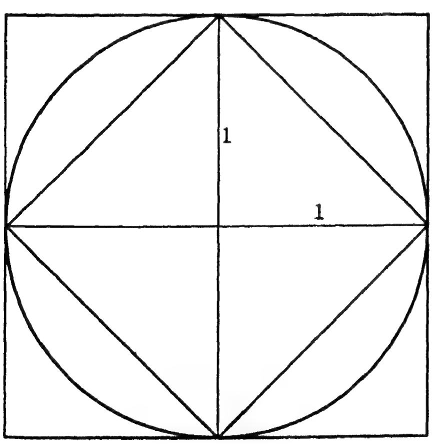


FIG. 146

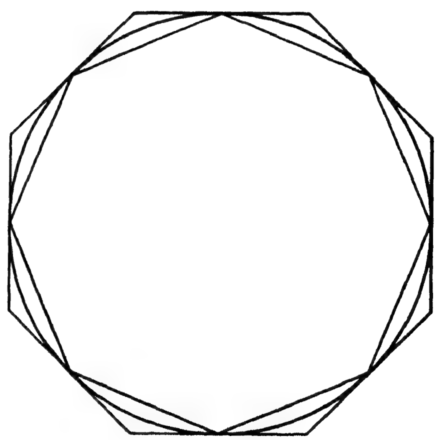


FIG. 147

that the area of the inscribed square is 2 and the area of the circumscribed square is 4. Thus $2 < \pi < 4$. This is a first approxima-

* I Kin. 7:23.

† The ancient artisan probably worked by trial and error, making no use of theoretical computations. But modern industry could hardly do without a good approximation to π.

tion. Now we can double the number of sides. It is not hard to calculate by the methods of elementary geometry that the area of the circumscribed regular 8-gon is $3.3137 \cdots$, and the area of the inscribed regular 8-gon is $2.8284 \cdots$. Thus $2.8284 \cdots < \pi < 3.3137 \cdots$. If we continue to increase the number of sides, we get the results in the following table:

No. of sides	*Area of inscr. reg. polygon (approx.)*	*Area of circum. reg. polygon (approx.)*	*Approximation of* π
4	2	4	
8	2.8284	3.3137	
16	3.0615	3.1826	3. +
32	3.1214	3.1517	3.1 +
64	3.1365	3.1441	
128	3.1403	3.1422	3.14 +
256	3.1413	3.1418	3.141 +

As we increase the number of sides, the area of the inscribed (or circumscribed) regular polygon approaches the area of the circle as a limit. In fact, the area of the circle may be defined as the limit of the area of the inscribed regular n-gon as n increases indefinitely. By taking n large enough we can approximate π as closely as we please. Archimedes carried this kind of process far enough to ascertain that $3\frac{10}{71} < \pi < 3\frac{1}{7}$, using a polygon of 96 sides. Archimedes' method is called the method of exhaustion, not because of what happens to its user, but because the growing inscribed regular polygon gradually exhausts the area of the circle. This method resembles and, indeed, contains the germ of the modern subject called integral calculus, which we shall take up briefly in the next sections. The idea of approximating a curved area by means of polygonal areas will be found again in modern form when we discuss the definite integral in section 114.

111. Anti-derivatives. If a function has a constant value for all values of x then its derivative is equal to zero. Conversely, it

is plausible and true that if the derivative of a function is zero for all values of x, then the function is a constant, although we shall not prove this here. Thus if two functions $F(x)$ and $G(x)$ have the same derivative at every value of x, (that is, $F'(x) = G'(x)$ for every value of x) then their difference $G(x) - F(x) = C$, or $G(x) = F(x) + C$, where C is some constant, because the derivative of $G(x) - F(x)$ is $G'(x) - F'(x) = 0$ by hypothesis. Conversely, if $G(x) = F(x) + C$ where C is any constant, then we obtain $G'(x) = F'(x)$ by differentiating both sides, since the derivative of C is zero. We have the following theorem.

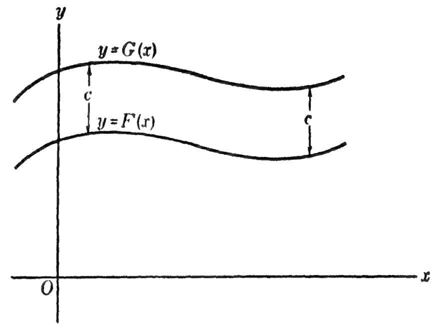


Fig. 148

Theorem 1. *If two functions $F(x)$ and $G(x)$ have the same derivative, then $G(x) = F(x) + C$ where C is some constant, and conversely.*

Geometrically, this means that two curves $y = F(x)$ and $y = G(x)$ have parallel tangents for each value of x if and only if the ordinates of the two curves for each value of x have a constant difference (Fig. 148).

Definition. *If the derivative of $F(x)$ is $f(x)$ then $F(x)$ is called an* ***anti-derivative*** *of $f(x)$.*

By the above theorem, all the anti-derivatives of $f(x)$ are given by the expression $F(x) + C$ where C may be any constant whatever. Thus an anti-derivative of $f(x)$ is not uniquely determined but if $F(x)$ is any one of them, any other can be obtained by adding a suitable constant to $F(x)$. The expression $F(x) + C$, which represents all the anti-derivatives is called the **indefinite integral** of $f(x)$ and is often written as $\int f(x)\,dx$ (read, the indefinite integral of $f(x)$ with respect to x) for a reason which will appear shortly. The process of finding the anti-derivatives of a function is called **integrating** the function. By definition, we have

$$\int f(x)\,dx = F(x) + C \text{ if and only if } F'(x) = f(x).$$

To integrate a polynomial is an easy matter if we recall the results of differentiation. For to ask for the indefinite integral of

$f(x)$ is to ask, "What functions have $f(x)$ as their derivative?" That is, by definition, $\int f(x)\,dx = F(x) + C$, where C is any constant, if and only if $\frac{d}{dx}F(x) = f(x)$. For example $\int x^3\,dx = \frac{x^4}{4} + C$ since $\frac{d}{dx}\left(\frac{x^4}{4}\right) = x^3$. In general, if we differentiate $\frac{x^{n+1}}{n+1}$ we obtain $\frac{n+1}{n+1}x^n$ or x^n, where n is any natural number; hence,

$$(1) \qquad \int x^n\,dx = \frac{x^{n+1}}{n+1} + C \textit{ if } n \textit{ is any natural number.}$$

Similarly, since $\frac{d}{dx}(kx) = k$, where k is a constant, we have

$$(2) \qquad \int k\,dx = kx + C.$$

We shall also use the following two theorems without proof:

$$(3) \qquad \int [f(x) + g(x)]\,dx = \int f(x)\,dx + \int g(x)\,dx,$$

$$(4) \qquad \int kf(x)\,dx = k\int f(x)\,dx, \textit{ where } k \textit{ is any constant.}$$

By means of (1), (2), (3), and (4), we can integrate any polynomial, as in the example below.

Integration (in this sense of finding the indefinite integral) and differentiation are inverse processes, much as multiplication and factoring are inverse processes. You recall that factoring was done essentially by remembering the results of multiplication.

Example.
$$\begin{aligned}
&\int (4x^3 + 3x + 4)\,dx \\
&= \int 4x^3\,dx + \int 3x\,dx + \int 4\,dx && \text{by (3)} \\
&= 4\int x^3\,dx + 3\int x\,dx + \int 4\,dx && \text{by (4)} \\
&= 4\cdot\frac{x^4}{4} + C_1 + 3\frac{x^2}{2} + C_2 + 4x + C_3 && \text{by (1) and (2)} \\
&= x^4 + \frac{3x^2}{2} + 4x + C,
\end{aligned}$$

denoting the sum of the constants $C_1 + C_2 + C_3$ by the single letter C. To check, we may differentiate our answer, obtaining the function we integrated, just as we checked factoring by multiplying the factors to obtain the original expression.

EXERCISES

Find the indefinite integral of each of the following functions, and check:

1. $x + 3.$ **2.** $3x^2 - x + 4.$ **3.** $5.$ **4.** $6x^2 + 4x + 1.$
5. $9 - 6x^2.$ **6.** $ax^2 + bx + c.$ **7.** $ax + b.$ **8.** $3x - 2.$
9. $96 - 32t.$

112. The constant of integration. The additive constant C which occurs in the indefinite integral may often be determined by the conditions of the problem.

Example 1. The slope of the tangent to a certain curve at any point (x,y) is equal to $x/2$. The curve passes through the point $(2,4)$. Find the equation of the curve.

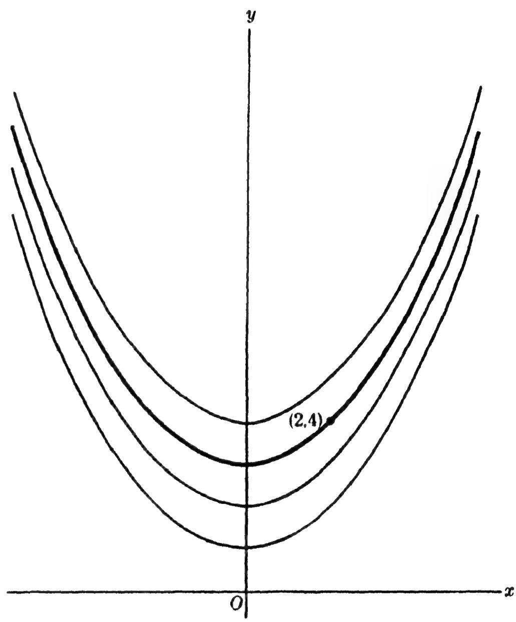


FIG. 149

Let $y = F(x)$ be the equation of the curve. We have to find the function $F(x)$. The first sentence tells us that $F'(x) = \frac{dy}{dx} = x/2$. Since the derivative of $F(x)$ is $x/2$, $F(x)$ is an anti-derivative of $x/2$. Hence, integrating, we obtain $y = \frac{1}{4}x^2 + C$. The second sentence tells us that at $x = 2$, $y = 4$. Hence $4 = \frac{1}{4}2^2 + C$ or $C = 3$. Thus $y = \frac{1}{4}x^2 + 3$ is the equation of the curve (Fig. 149). Other values of C would yield curves with parallel tangents, but not passing through the desired point.

Example 2. It is found by experiment that if a projectile is thrown directly upward, its acceleration is -32 ft. per sec. per sec. Let v denote the velocity of the projectile. By definition, acceleration $a = \frac{dv}{dt}$ or the rate of change of the velocity with

respect to time. Hence $\frac{dv}{dt} = -32$. Integrating we get $v = -32t + C$. Thus, at $t = 0$, $v = C$ or C is the initial velocity; call it v_0. Thus,

(1) $$v = -32t + v_0.$$

Now if h is the height of the projectile, $v = \frac{dh}{dt}$ or the rate of change of the height with respect to time. Integrating (1) we get $h = -16t^2 + v_0t + C_0$. At $t = 0$, $h = C_0$, or C_0 is the initial height; call it h_0. Hence

(2) $$h = -16t^2 + v_0t + h_0$$

gives the height as a function of t. For example, if a projectile is thrown straight up from a roof 30 ft. high with an initial velocity of 100 ft. per sec., then its height, measured in feet, at the end of t seconds is given by $h = -16t^2 + 100t + 30$. Its velocity at the end of t seconds is given by $v = -32t + 100$.

EXERCISES

Find the equation of the curve whose tangent has the slope:

1. 3 at all values of x and passes through the point (2,7).

2. m at all values of x and passes through the point $(0,p)$.

3. $2x$ at all values of x and passes through the point (3,5).

4. $3x^2 + 2x + 3$ at all values of x and passes through the point (0,1).

5. $2x - 3$ at all values of x and passes through the point (1,3).

6. Find the function $h = F(t)$ whose derivative with respect to t is $128 - 32t$ and for which $h = 6$ when $t = 0$.

7. The acceleration of a falling body is 32 ft. per sec. per sec. With what velocity will a stone strike the ground if it is dropped from a roof 100 ft. high?

8. If a stone is thrown straight down with an initial velocity of 60 ft. per sec., from a roof 100 ft. high, with what velocity will it strike the ground?

9. A projectile is thrown straight up from the edge of a roof 64 ft. high with an initial velocity of 48 ft. per sec. With what velocity will it strike the ground?

113. The area under a curve. Consider the curve $y = f(x)$ where $f(x)$ is a continuous single-valued function. Suppose the curve lies above the x-axis, for simplicity. By the **area under the curve between a and b** we mean the area enclosed by the curve, the x-axis, and the vertical lines $x = a$ and $x = b$. Thus $MNCD$ is the area A under the curve $y = f(x)$ between

x_0 and x (Fig. 150). Here A is a function of x. Let us find the rate of change of A with respect to x. Take a point E with the coordinates $(x + \Delta x, 0)$. The change in area ΔA induced by this change Δx in x is the area $CKGED$. Note that $CD = f(x)$ and $EG = f(x + \Delta x)$. Now the area $CKGED$ is more than $CFED = f(x) \cdot \Delta x$ and less than $HGED = f(x + \Delta x) \cdot \Delta x$. Hence it is equal to Δx multiplied by some value $\bar{y}$ between $y = f(x)$ and $y + \Delta y = f(x + \Delta x)$, such as the length of KL; that is, $\Delta A = \bar{y} \cdot \Delta x$ where $y < \bar{y} < y + \Delta y$. Hence $\dfrac{\Delta A}{\Delta x} = \bar{y}$. As $\Delta x \to 0$, so does $\Delta y \to 0$ and $\bar{y} \to y$. Hence $\dfrac{dA}{dx} = y$, or $\dfrac{dA}{dx} = f(x)$. Therefore

$$A = \int f(x)\,dx + C,$$

where C is some constant.

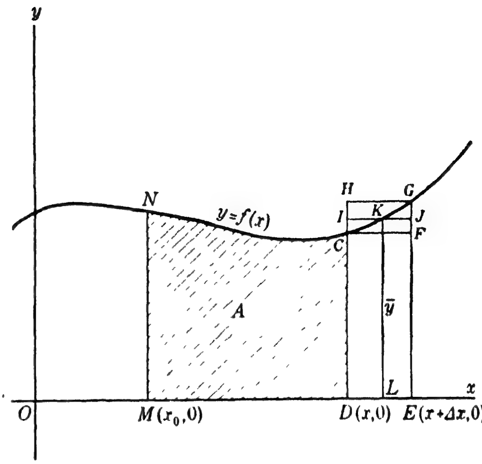


FIG. 150

Intuitively one may think of the area under the curve between $x = a$ and $x = b$ as being swept out by a vertical line-segment of varying length moving from the position $x = a$ to $x = b$.

Example. Find the area under the parabola $y = \frac{1}{4}x^2 + 1$ between $x = 1$ and $x = 3$. We have $A = \int\left(\frac{1}{4}x^2 + 1\right)dx = \frac{x^3}{12} + x + C$. Now at $x = 1$, the area swept out is zero, since the "sweeping out" process is to start at $x = 1$. Hence $0 = \frac{1}{12} + 1 + C$ or $C = -\frac{13}{12}$. Therefore $A = \frac{x^3}{12} + x - \frac{13}{12}$. Hence

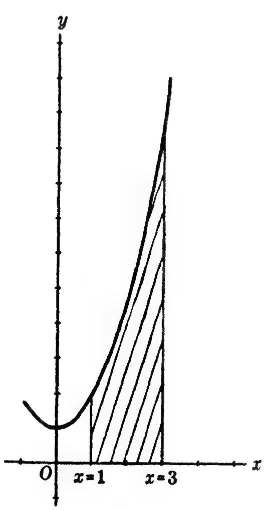


FIG. 151

at $x = 3$, $A = 27/12 + 3 - 13/12 = 50/12 = 25/6$, which is the required area.

EXERCISES

Find the area under the curve:

1. $y = 3x + 4$ from $x = 1$ to $x = 4$. Verify your answer by elementary geometry.

2. $y = 15 - 2x$ from $x = 1$ to $x = 4$. Verify your answer by elementary geometry.

3. $y = \frac{1}{2}x^2$ from $x = 1$ to $x = 4$.

4. $y = x^2 - x + 10$ from $x = 2$ to $x = 5$.

5. $y = 10 - x^2$ from $x = 0$ to $x = 3$.

6. (a) $y = 3x^2 + 2x$ from $x = 0$ to $x = 2$.
(b) $y = 4x^3 + 3$ from $x = 0$ to $x = 2$.

114. The definite integral. The area A under a curve $y = f(x)$ can be considered in another way. Consider the area under $y = f(x)$ from $x = a$ to $x = b$. Divide the interval from a to b on the x-axis into n equal intervals of length Δx. Therefore, $n \cdot \Delta x = b - a$. Let y_1, y_2, $\cdots$, y_n be the values of y at the beginning of each of these small intervals respectively. Then the area of the "staircase" (Fig. 152) is

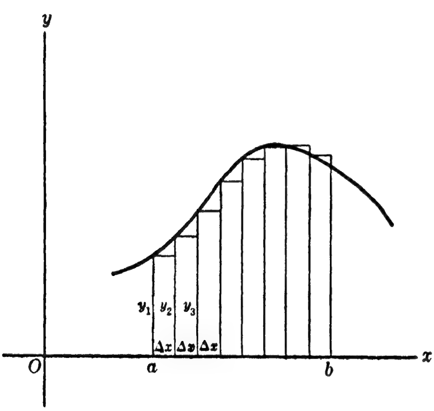


FIG. 152

$$S = y_1 \cdot \Delta x + y_2 \cdot \Delta x + \cdots + y_n \cdot \Delta x.$$

Now this sum approximates the area under the curve if Δx is small enough. In fact, if we make Δx smaller and smaller, allowing n to increase correspondingly, the sum S approaches the area A as a limit. Note that since $\Delta x = \frac{1}{n}(b - a)$, we must have $\Delta x \to 0$ when n increases indefinitely. We may define the area A as

$$A = \lim_{\substack{\Delta x \to 0 \\ n \uparrow}} (y_1 \cdot \Delta x + y_2 \cdot \Delta x + \cdots + y_n \cdot \Delta x).$$

This limit of a sum is called the **definite integral of $f(x)$ from a to b,** and is denoted symbolically by

$$\int_a^b f(x)\, dx.$$

The numbers a and b are called the **limits of integration.** The integral sign $\int$ comes from the idea of the definite integral as the limit of a sum; the sign is nothing but an elongated S.

The definite integral defined here has no a priori connection with the indefinite integral defined in section 111. But it is possible to prove the following remarkable theorem known as the *fundamental theorem of integral calculus:*

If $F(x)$ is any anti-derivative of $f(x)$, then

$$\int_a^b f(x)\, dx = F(b) - F(a).$$

This theorem may be used conveniently to calculate areas under curves as in the following example already studied in section 113.

Example. Find the area under the parabola $y = \frac{1}{4}x^2 + 1$ between $x = 1$ and $x = 3$. The area desired is $\int_1^3 \left(\frac{1}{4}x^2 + 1\right) dx$. Now the indefinite integral $\int \left(\frac{1}{4}x^2 + 1\right) dx = \frac{x^3}{12} + x + C$. Hence the function $F(x) = \frac{x^3}{12} + x$ is an anti-derivative of $x^2 + 1$. By the theorem above,

$$\int_1^3 \left(\frac{1}{4}x^2 + 1\right) dx = F(3) - F(1) = \frac{3^3}{12} + 3 - \left(\frac{1^3}{12} + 1\right) = 25/6.$$

EXERCISES

Do each of the exercises of section 113 again, this time using the fundamental theorem of integral calculus as in the above example.

The idea of the definite integral as the limit of a sum of small quantities is the central idea of the **integral calculus** and is essential in many connections.

For example, the length of a curve can be defined as the limit of the length of an inscribed broken line; that is, the limit of a sum of chords as the number of chords becomes larger and the length of each chord becomes smaller (Fig. 153). Note that one cannot lightly pass off the definition of the length of a curve by saying that it is the "number of inches" in the curve. For to find the "number of inches" in a straight line we lay off a straight ruler along it; but no part of a curved line can be made to coincide exactly with a straight ruler. To speak of the length of a curved line as being the length of the straight line obtained by "bending" it until it is straight is equally futile. For you would want to bend it without stretching or contracting, which means that you want to bend it so that its "length" is unchanged during the process. But this presupposes that "length" has already been defined, and would therefore be a circular definition.

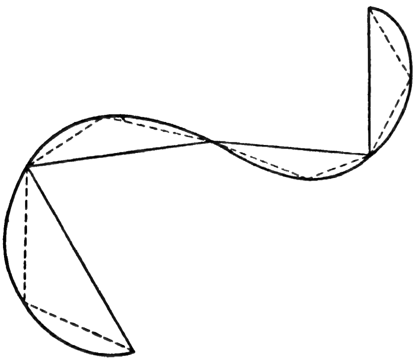


FIG. 153

Similarly, the volume of a curved solid may be defined as the limit of a sum of the volumes of thin slices as the number of slices becomes larger and each slice thinner (Fig. 154). These and many other geometrical and physical concepts may be defined as definite integrals and studied by means of the integral calculus.

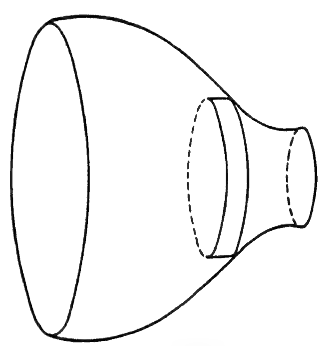


FIG. 154

115. Infinite series. The idea of limit enters elementary mathematics in another important way. Suppose we were to write the expression

(1) $$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\cdots+\frac{1}{2^n}+\cdots$$

where the three dots at the end indicate that the expression goes on forever. Such an expression is called an **infinite series.** What

can this possibly mean? Surely it cannot mean that we are expected to spend the rest of our lives adding the successive terms to the partial sums already obtained. For, having thus misspent our lives we should obviously have to leave the task unfinished and curse our descendants with the burden of continuing; this little example might prove to be the ruin of the human race. What we do mean by the expression (1) is the following. Let s_n be the sum of the first n terms, where n is any natural number. Thus

$s_1 = \frac{1}{2}$, $s_2 = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$, $s_3 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}$, $s_4 = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16}$, $\cdots$; in general, we could prove that $s_n = \frac{2^n - 1}{2^n}$. If the sequence $s_1, s_2, \cdots, s_n, \cdots$ has a limit S we call S the *sum*

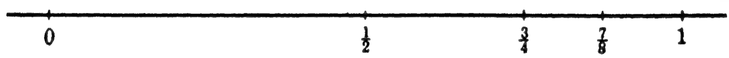


FIG. 155

of the infinite series (1). In this example $\lim_{n\uparrow} s_n = 1$ and hence we write

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n} + \cdots = 1.$$

The meaning of this statement involves the idea of limit. No matter how large a natural number n we take, the sum s_n of the first n terms is never equal to 1, but as n increases indefinitely, s_n approaches 1 as a limit. In general, given an infinite series

$$a_1 + a_2 + \cdots + a_n + \cdots$$

we say that the series has the **sum** S if the sequence of the partial sums

$$\begin{aligned}
s_1 &= a_1 \\
s_2 &= a_1 + a_2 \\
s_3 &= a_1 + a_2 + a_3 \\
&\cdot \\
&\cdot \\
&\cdot \\
s_n &= a_1 + \cdots + a_n \\
&\cdot \\
&\cdot \\
&\cdot
\end{aligned}$$

approaches S as a limit. If this happens the series is said to **converge;** if not, it is said to **diverge.**

Infinite series are of the greatest practical value in mathematics and its applications. For example, if one wishes to calculate a good approximate value for a function which is difficult to compute directly, it is often possible to express the function as an infinite series and to approximate it by taking the sum of the first few terms. If a better approximation is desired, one can take the sum of a few more terms. This method is actually employed by mathematicians in the construction of tables, for example, of log x. Thus it can be shown that

$$\text{(2)} \qquad \log_e 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} + \cdots + \frac{(-1)^{n+1}}{n} + \cdots$$

$$\text{(3)} \qquad e = 2 + \frac{1}{2} + \frac{1}{3 \cdot 2} + \frac{1}{4 \cdot 3 \cdot 2} + \frac{1}{5 \cdot 4 \cdot 3 \cdot 2} + \frac{1}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2} + \cdots$$

and

$$\text{(4)} \qquad \pi = 4 - \frac{4}{3} + \frac{4}{5} - \frac{4}{7} + \frac{4}{9} - \frac{4}{11} + \frac{4}{13} - \cdots.$$

One can in fact decide how many terms one has to take to make the approximation accurate to the desired number of decimal places. Series (4) approaches its limit π so slowly that one would have to take a great many terms to get even the first few places of π. It is consequently virtually useless for the purpose of computing π. A better expression for this purpose is

$$\text{(5)} \qquad \pi = 16\left[\frac{1}{5} - \frac{1}{3 \cdot 5^3} + \frac{1}{5 \cdot 5^5} - \frac{1}{7 \cdot 5^7} + - \cdots\right] - 4\left[\frac{1}{239} - \frac{1}{3 \cdot 239^3} + \frac{1}{5 \cdot 239^5} - + \cdots\right].$$

Besides the use of infinite series in the construction of tables of various sorts, they are of the greatest value to engineers, and others, in a great variety of practical situations. Their value in the development of pure mathematics is also immense.

The idea of limits disturbed the ancients, although a good logical treatment of them is of very recent date. For example, Zeno (5th century B.C.) argued as follows. If a man has to reach a

point a mile away, he cannot do it. For he first has to traverse the first half mile; then he would have to traverse the next quarter of a mile; then the next eighth of a mile, and so on. Thus he has to traverse an infinite succession of intervals. Zeno argued that to perform an infinite succession of acts should take an infinite length of time and therefore the man can never reach his goal. This "paradox" of Zeno has been the subject of much vague discussion. It is really not a paradox at all; one must simply realize that the "sum" of an infinite number of terms may very well be finite. In fact, if it takes the man 1 hour to walk a mile, he will walk the first half mile in a half hour, and so on. Therefore it will take him $\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^n} + \cdots$ hours to walk a mile. Despite the fact that this is an infinite series, the sum is one hour.

The same mathematical situation was presented by Zeno in the form of the "paradox" of Achilles and the tortoise. Achilles, who could run 10 yards in 1 second, was to run a race with a tortoise, who could run only 5 yards in 1 second, but the tortoise was to have a handicap of 10 yards. Zeno asserted that Achilles could never catch the tortoise for the following reasons. By the time Achilles had run the 10 yards from his starting point A_1 to the point T_1 where the tortoise started, the tortoise would have run ahead 5 yards to T_2; by the time Achilles had run the 5 yards to T_2 the tortoise would no longer be there, having run on the $2\frac{1}{2}$ yards to T_3; and so on. Here again we have an infinite series, for Achilles requires 1 second to go from A_1 to T_1, $\frac{1}{2}$ second to go from T_1 to T_2, $\frac{1}{4}$ second to go from T_2 to T_3, and so on. Hence the

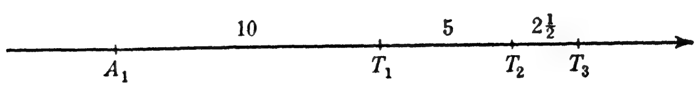


Fig. 156

time required by Achilles to catch the tortoise is the sum of the infinite series

$$1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^{n-1}} + \cdots \text{ which is 2 seconds.}$$

It should be said, in justice to Zeno, that he did not really believe anything so contrary to experience as the assertion that Achilles could not catch the tortoise. In fact, it is a simple problem in elementary algebra to find out how long it would take for Achilles to catch the tortoise. We leave this as an exercise for the student (see exercise 18 section 41). What Zeno was concerned about was locating the catch in his argument.

EXERCISES

1. Using the first five terms of series (2) calculate a four-place decimal approximation for $\log_e 2$.

2. Obtain a four-place decimal approximation for e from series (3) using (*a*) the first 4 terms; (*b*) the first five terms; (*c*) the first six terms.

3. Obtain a four-place decimal approximation for π from (5) using the first two terms in each bracket; (*b*) the first three terms in each bracket.

116. Conclusion. Our discussion of calculus has been very sketchy and largely intuitive because of the technical difficulty of discussing the subject logically. A strictly logical treatment is given in more advanced courses. Our purpose here was to introduce you to its most elementary ideas and to give you some hint of its practical importance and its power in applications to geometry and physics. The far reaching effects of the calculus in astronomy, physics, engineering, etc., can not be overemphasized. The calculus initiated a new era and may be said to mark a turning point in the history of mathematics and its applications.

REFERENCES

E. T. Bell, *The Handmaiden of the Sciences*, Williams & Wilkins, Baltimore, 1937.

E. T. Bell, *Men of Mathematics*, Simon & Schuster, N. Y., 1937. (Biographies of Newton, Leibniz, et al.)

E. T. Bell, *The Development of Mathematics*, McGraw-Hill, N. Y., 1940.

G. Berkeley, *The Analyst*, London, 1734.

R. Courant, *Differential and Integral Calculus*, (translated by E. J. McShane), Nordemann, N. Y., 1937. (For the serious student.)

A. Dresden, *Introduction to the Calculus*, Henry Holt, N. Y., 1940. (For the serious student.)

A. Dresden, *Invitation to Mathematics*, Henry Holt, N. Y., 1936.

F. L. Griffin, *Introduction to Mathematical Analysis*, Houghton Mifflin, Boston, 1921.

S. P. Thompson, *Calculus Made Easy*, Macmillan, London, 1910.
A. N. Whitehead, *Introduction to Mathematics*, Henry Holt, N. Y., 1911.
Histories of Mathematics, as before.
Text-books on Differential and Integral Calculus.

Chapter XII

TRIGONOMETRIC FUNCTIONS

117. Introduction. The word "trigonometry" means the study of the measurement of triangles. The student will doubtless recall that much of his high school course in plane geometry was devoted to the study of triangles. Why is the triangle given so much attention? It is easy to see that the triangle is the simplest possible figure, bounded by straight line segments, which is capable of enclosing an area; that is, the triangle is the simplest possible polygon. This would explain why it is studied early in geometry. But more important than its simplicity is the fact that every polygon can be split up into triangles by drawing suitable diagonals (Fig. 157). That is, the triangles are fundamental building blocks out of which we can make any polygon whatever. Or, to put it another way, we can measure any polygonal figure by splitting it into triangles and measuring each triangle. Therefore the study of surveying polygonal plots of land, etc., is reduced to the study of the triangle.

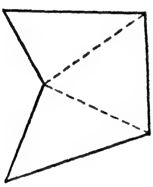


FIG. 157

An important class of functions, called the trigonometric functions, arises from the study of the measurement of triangles. Among their numerous applications, some of which will be mentioned below, is the remarkable achievement of the measurement of inaccessible distances, such as the heights of mountains, the distance from the earth to the moon, or the distance from a cannon to an enemy supply base; all of these are distances which it is either impossible or inadvisable to measure directly by the application of a tape measure. Their most elementary aspects go back to ancient times but many of their most interesting phases are of recent origin. We shall discuss both the elementary applications to surveying and the measurement of inaccessible distances, and the more advanced applications to modern science.

118. Angular measure. Consider a line (extending indefinitely in both directions as opposed to a line-segment which has a definite length). Any point P on the line divides it into two parts, one on each "side" of the point. The set of all points on

FIG. 158

one side of the point P is called a **ray**. Thus P divides the line into two rays, PQ and PQ' (Fig. 158). The point P is called the **vertex** of each of these rays. The word "ray" is obviously suggested by light rays emanating from a source P. An **angle** is the figure consisting of (or, the set of points on) two rays having the same vertex (Fig. 159). We wish to associate with each angle a number called the measure of the angle. Intuitively, we want the measure of an angle to describe the smallest amount of rotation by which we can turn from the direction of one ray to the direction of the other. Perhaps the most obvious unit with which to measure the amount of rotation is a complete revolution; that is, the smallest actual rotation which brings you back to your initial direction (Fig. 160). But this unit is so large that it is cumbersome. It is more usual to choose as the unit of angular measure the **degree,** which may be defined as one 360th of a complete revolution. Thus we associate with each angle a definite number called the **measure** of the angle, namely the smallest number of degrees required to rotate from the direction of one ray of the angle to the direction of the other ray. For example, the angle in Fig. 161 has a measure of 90° or 1/4 of a complete revolution. The measure of an angle is a single-valued function of the angle, for to each angle there corresponds a unique number called its measure. *This function ranges between 0° and 180° only.*

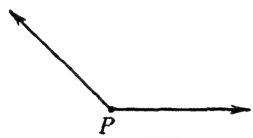


FIG. 159

FIG. 160

One sixtieth of a degree is called a **minute,** and one sixtieth of a minute is called a **second.** The symbol 5° 3′ 4″ means 5 degrees, 3 minutes, and 4 seconds.

FIG. 161

The **sexagesimal** or **degree system** of angular measure may have been used as far back as the time of the Babylonians (about two thousand years B.C. or perhaps earlier). It is sometimes said that they divided the circle into 360 equal parts because their year was taken as 360 days, but this is not certain. They probably knew that a chord equal in length to the radius could be laid off around the circle six times, thus inscribing a regular hexagon (Fig. 162), and they may have divided the central angle C into 60 equal parts because their notation for numbers was based on 60.

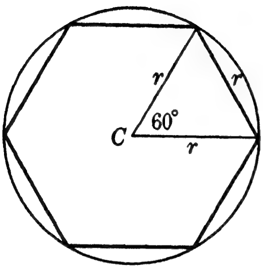


FIG. 162

Another unit of angular measure was introduced in modern times for theoretical reasons (see Remark 2, section 121). If an *arc* equal in length to the radius is laid off along the circumference of a circle, the angle formed at the center by the radii drawn to the extremities of this arc is uniquely determined; that is, if we did the same thing with any other circle, we would get an equal angle (angle O = angle O' in Fig. 163). The smallest rotation by which we can turn from the direction of one ray of this angle to the direction of the other ray is called a **radian.** Measuring an angle according to the number of radians it contains is very convenient for certain purposes. Since the circumference of a circle is given by $C = 2\pi r$, the length of a semi-circular arc is πr where r is the radius of the circle and $\pi = 3.14159265 \cdots$. Therefore an angular measure of 180° is equal to an angular measure of π radians; or

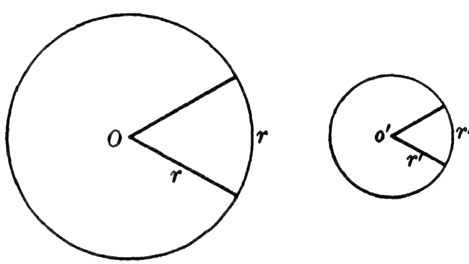


FIG. 163

(1) $$180° = \pi \text{ radians.}$$

Dividing both sides of (1) by 180 we get

$$1° = \pi/180 \text{ radians} = .01745 \cdots \text{ radians;}$$

dividing both sides of (1) instead by π we get

$$1 \text{ radian} = (180/\pi)° = 57.295 \cdots °.$$

This enables us to convert an angular measure from degrees to radians and conversely; just as the knowledge that 1 ft. = 1/3 yd. and 1 yd. = 3 ft. enables us to convert a measure of length from feet to yards and conversely. It is customary to use the symbol ° for degrees and to use no symbol at all when radians are meant. Thus an angular measure of 3° means 3 degrees while an angular measure of 3 means 3 radians.

It is customary to speak of an "angle" of 3° or an "angle" of 3 (radians) instead of an angle whose measure is 3° or 3. This abbreviated manner of speaking does no harm. Of course, the measure of an angle is only one of many functions that can be associated with an angle.

EXERCISES

Find the number of radians in an angle of:

1. 90°. **2.** 45°. **3.** 60°. **4.** 30°. **5.** 10°. **6.** 2°.

Find the number of degrees in an angle of:

7. π radians. **8.** $\pi/2$. **9.** $\pi/6$. **10.** $\pi/18$. **11.** 1.5.
12. 3. **13.** .5.

119. Similar triangles. Two triangles ABC and $A'B'C'$ are called **similar** if their angles are respectively equal, that is, if $A = A'$, $B = B'$, and $C = C'$. Denoting the length of the side opposite angle A by a, and so on as in Fig. 164, we must recall the following important theorems of elementary geometry.

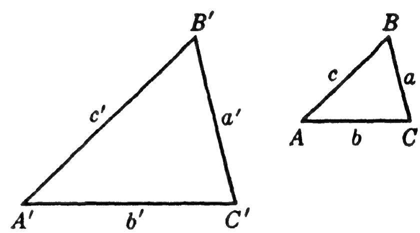


FIG. 164

THEOREM 1. *If triangles ABC and $A'B'C'$ are similar then* $\frac{a}{b} = \frac{a'}{b'}, \frac{b}{c} = \frac{b'}{c'}, \frac{a}{c} = \frac{a'}{c'}$.

THEOREM 2. *The sum of the angles of any triangle is 180°.*

COROLLARY. *If two angles of one triangle are respectively equal to two angles of another triangle, then the triangles are similar.* (Why?)

Theorem 1 can be used to measure inaccessible distances. For example, the height of a vertical greased pole may be obtained

as follows. Place a short stick perpendicular to the ground. Measure it, its shadow, and the shadow of the pole. Thus (Fig. 165) BC, AB, and $A'B'$, are known quantities. But since nearby sun's rays are parallel, for all practical purposes, angle A = angle A' and the triangles are therefore similar. (Why?) Hence $\frac{B'C'}{A'B'} = \frac{BC}{AB}$ or $B'C' = \frac{BC}{AB} \cdot A'B'$. For example, if the stick BC were 5 ft. long, the stick's shadow AB were 3 ft. long, and the pole's shadow $A'B'$ were 30 ft. long, then the length $B'C'$ of the pole would be $\frac{5}{3} \cdot 30 = 50$ ft. Notice that we have not had to climb the greased pole with a tape measure. This method was probably used by Thales (about 600 B.C.), one of the earliest great Greek mathematicians, to calculate the height of an Egyptian pyramid, with the aid of his walking stick, to the consternation of the Egyptian priests.

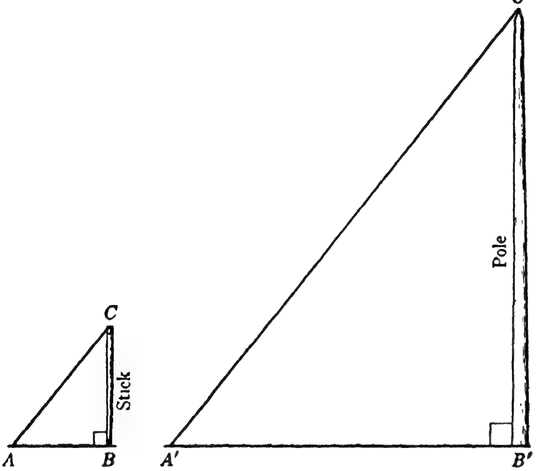


FIG. 165

Exercise. If a 5 ft. pole, at right angles to the ground, casts a 3 ft. shadow at the time when the shadow of an apartment house is 48 ft. long, how high is the house?

120. The trigonometric functions of acute angles. An angle whose measure is greater than 0° but less than 90° is called **acute.** Consider a right triangle ABC with the right angle at C, lettered in the standard way, as in Fig. 166. Side c is called the **hypotenuse,** a is called **opposite to A** and **adjacent to B,** b is called **opposite to B** and **adjacent to A.** Any ratio of the lengths of two sides of the triangle, for example a/c, is a number which is completely determined by angle A alone and is independent of the size of the triangle. That is, if we took a larger or smaller right triangle $A'B'C'$ with $A = A'$ then $a'/c' = a/c$ since these triangles

would be similar. Thus the number a/c is a function of the angle A; it is called the **sine** of A. Similarly any other ratio of two of the sides depends only on the measure of angle A. There are six possible ratios that can be formed from the three sides of a right triangle. We give each of them a name by adopting the following definitions.

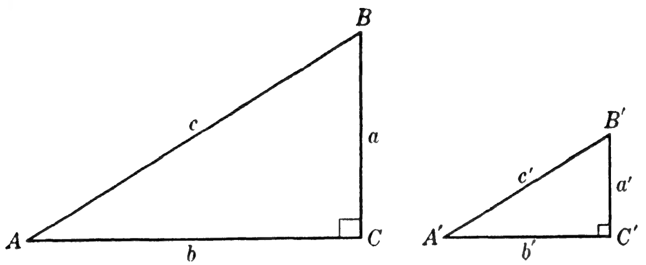


FIG. 166

DEFINITIONS.

$$\begin{aligned} \sin A &= a/c = opp./hyp. & \csc A &= c/a = hyp./opp. \\ \cos A &= b/c = adj./hyp. & \sec A &= c/b = hyp./adj. \\ \tan A &= a/b = opp./adj. & \cot A &= b/a = adj./opp. \end{aligned}$$

The abbreviations stand for **sine, cosine, tangent, cosecant, secant, cotangent,** of A, respectively. These queerly named functions of angle A are called **trigonometric functions** or **trigonometric ratios.** The origin of the names of the trigonometric functions will be discussed in section 125.

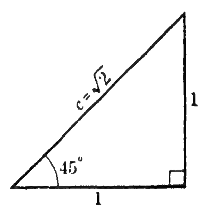


FIG. 167

Example 1. Consider an isosceles right triangle whose leg is one unit long. From the Pythagorean theorem we find that $c^2 = 1^2 + 1^2$ or $c^2 = 2$; hence the hypotenuse is equal to $\sqrt{2}$. Thus $\sin 45° = \dfrac{1}{\sqrt{2}} = \dfrac{1}{\sqrt{2}} \cdot \dfrac{\sqrt{2}}{\sqrt{2}} = \dfrac{\sqrt{2}}{2} = \dfrac{1.414 \cdots}{2} =$ $.707 \cdots$ approximately,* while $\tan 45° = 1$. If we took a differ-

* The process by which we obtained $\sqrt{2}/2$ from $1/\sqrt{2}$ is called **rationalizing the denominator.** It was done *for the purpose of obtaining a decimal expression,* since it is easier to divide $1.414 \cdots$ by 2 than to divide 1 by $1.414 \cdots$. For most purposes other than decimal evaluation, there is no reason to regard $\sqrt{2}/2$ as being "simpler" than $1/\sqrt{2}$. In what follows, we shall seldom rationalize denominators.

ent right triangle whose leg was 10 units long, the hypotenuse would have been $10\sqrt{2}$ and we would obtain the same values for sin 45°, etc. (Why?)

Example 2. Consider the equilateral triangle ABD whose side is 2 units long, bisected as in Fig. 168. Triangle ABC is called a 30°–60°–90° triangle. Clearly $AC = 1$ and from the Pythagorean theorem we find that $1^2 + a^2 = 2^2$ or $a^2 = 3$; hence $a = BC = \sqrt{3}$; hence $\sin 60° = \sqrt{3}/2 = 1.732 \cdots /2 = .866 \cdots$ approximately, $\sin 30° = 1/2 = .500$, etc.

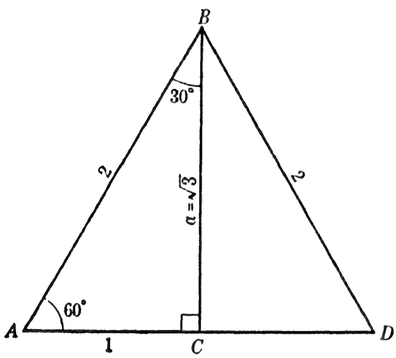


Fig. 168

Two different acute angles cannot have equal trigonometric functions. For suppose $\sin A = \sin A'$; that is $a/c = a'/c'$. Then it follows from a theorem of geometry that the two right triangles are similar and hence $A = A'$. Thus *an acute angle is uniquely determined if the value of one of its trigonometric functions is given.* In particular, given one trigonometric function of an acute angle, we can find the other five.

Example 3. Given the acute angle A with $\sin A = 5/13$, find $\cos A$ (Fig. 169).

Since $\sin A = 5/13$ we may place the angle A in a right triangle with $a = 5$ and $c = 13$. Then $5^2 + b^2 = 13^2$ or $b^2 = 169 - 25 = 144$. Hence $b = 12$. Therefore $\cos A = 12/13$.

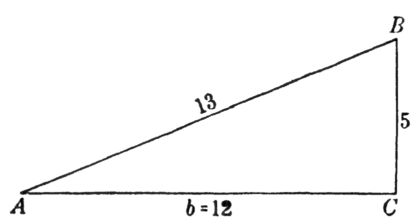


Fig. 169

If two trigonometric functions are such that the name of one can be obtained from the name of the other by prefixing or omitting the prefix "co-," we shall call them **co-functions.** For example, sine and cosine are co-functions. Let A and B be two **complementary** angles (that is, $A + B = 90°$) in a right triangle (Fig. 166). Since the side opposite A is adjacent to B, and vice

versa, we have $\sin A = a/c = \cos B$, $\cos A = b/c = \sin B$, and so on. In general, we have the following theorem.

THEOREM. *Any trigonometric function of an acute angle is equal to the co-function of its complement.*

Writing B as $90° - A$, this may be written as $\sin A = \cos (90° - A)$, $\tan A = \cot (90° - A)$, and so on.

EXERCISES

Find the values of the six trigonometric functions of an angle, in a right triangle, whose:

1. opposite side is 3 and hypotenuse 5.

2. adjacent side is 5 and hypotenuse 13.

3. adjacent side is 12 and opposite 5.

4. adjacent side is 3 and opposite 4.

5. adjacent side is 5 and opposite 4.

6. adjacent side is 3 and hypotenuse 4.

7. adjacent side is 4 and opposite $\sqrt{3}$.

Find the values of the six trigonometric functions of:

8. 45°. **9.** 30°. **10.** 60°.

Find the number of degrees in the acute angle A if:

11. $\sin A = 1/\sqrt{2}$. **12.** $\tan A = \sqrt{3}$. **13.** $\cos A = 1/2$.

14. $\tan A = 1$. **15.** $\cos A = \sqrt{3}/2$. **16.** $\sin A = \sqrt{3}/2$.

Find all the trigonometric functions of the acute angle A given that:

17. $\sin A = 6/10$. **18.** $\sin A = 3/4$. **19.** $\cos A = 5/13$.

20. $\sin A = 5/13$. **21.** $\cos A = 3/4$. **22.** $\tan A = 4/5$.

23. $\tan A = 4/3$. **24.** $\cos A = 5/12$. **25.** $\tan A = 5/3$.

26. $\csc A = 5/3$. **27.** $\sec A = 10/8$. **28.** $\cot A = 5/6$.

29. $\sin A = .8$. **30.** $\cos A = .7$. **31.** $\tan A = 1.2$.

32. $\csc A = 1.3$. **33.** $\sec A = 1.4$. **34.** $\cot A = 1.5$.

35. If A and B are complementary angles and $\sin A = 1/3$, find (*a*) $\cos B$; (*b*) $\cos A$; (*c*) $\sin B$.

36. If A and B are complementary angles, and $\sin A = 3/5$, find all the trigonometric functions of both A and B.

37. If $\sin 27° = .454$, find $\cos 63°$.

38. If $\cot 41° = 1.15$, find $\tan 49°$.

39. If $\sec 44° = 1.39$, find $\csc 46°$.

121. Applications. Suppose we observe that the shadow cast by a pole is 30 ft. long at the instant when the **angle of elevation** of the sun (that is, the angle made by the sun's rays and the horizontal ground) is 38° (Fig. 170). How can we tell the height h of the pole? Clearly

$$h/30 = \tan 38°$$

so that if we knew the value of tan 38° we could get $h = 30 \cdot \tan 38°$ without more ado. Thales obtained the value of this ratio from his walking stick and its shadow (section 119). But we would need no walking stick at all if we had a table giving the values of the trigonometric functions (or ratios) for various angles. Thus we see the utility of having a table of the trigonometric functions for acute angles. Such a table is found at the end of the book. From it we find that tan 38° = .7813. Hence, $h = 23.44$ ft.

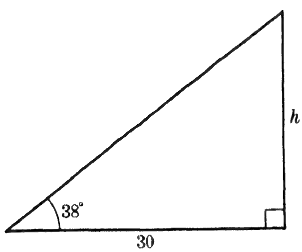


Fig. 170

One of the most important elementary applications of the trigonometric functions is to such problems of indirect measurement. Other applications are contained in the examples and exercises.

In reading the tables it must be understood that the headings at the top go with the angles at the left, while the headings at the bottom go with the angles at the right. This compact arrangement of the table is made possible by the theorem of section 120 on co-functions. For a glance at the table will show that sin 60°, for example, is located in the same place as cos 30°. This can be done because sin 60° = cos 30°, since 30° and 60° are complementary. Hence the trigonometric functions of angles of 1, 2, · · ·, 89 degrees can be arranged in 45 lines instead of 89 lines. Note that the line in the table giving the functions of 0° and 90° must be regarded as senseless at the moment because we have

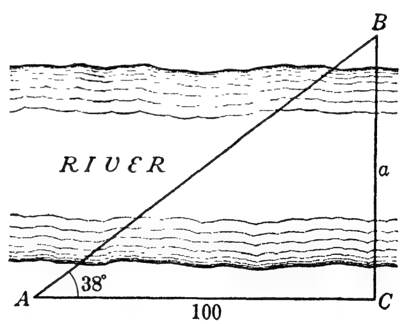
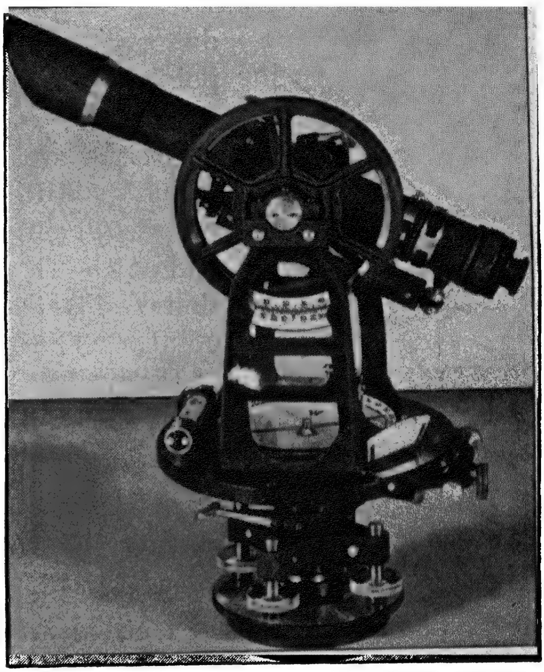


Fig. 171

not yet defined the trigonometric functions of angles, like 0° and 90°, which are not acute. These will be defined in section 124.

Example 1. Find the distance BC across the river in Fig. 171 if $AC = 100$ ft. and $A = 38°$ by direct observation.*

We are given the side adjacent to angle A and must find the side opposite angle A. Therefore we are concerned with the tangent of angle A. Clearly $a/100 = \tan 38°$. Hence $a = 78.13$ ft. Notice that this distance is determined from data which can be obtained while remaining on one side of the river, as follows. We can stand at C with a transit pointing toward some object at B and turn it through 90°. An assistant finds a place A directly in the line of sight of the turned transit and marks the spot. Then AC is measured. Marking C we move the transit to A, point it toward B, and turn it until we sight C. We then read off the angle A from the transit.



A transit

FIG. 172

Example 2. If a 30 ft. ladder is leaned against a window sill 20 ft. high, what angle will the ladder make with the ground?

We are given the side opposite angle A and the hypotenuse (Fig. 173). Therefore we are concerned with the sine of angle A. Clearly $\sin A = 20/30 = .6667$. From the table we find that the nearest value to .6667 in the sine column is .6691, which is $\sin 42°$. Hence we take $A = 42°$ as an approximate answer.

* In practice, angles are measured by means of a *transit*, which is really nothing more than a glorified protractor with a spy-glass to aid the vision. A transit mounted on a tripod may be seen on any empty lot just before a building goes up.

Example 3. If a projectile is fired at an angle A from the level of the ground its horizontal distance x and its height y at the end of t seconds are given by (Fig. 174)

(1) $$x = v_0 t \cos A$$

(2) $$y = v_0 t \sin A - 16\, t^2$$

where v_0 is the initial velocity and distances are measured in feet. Suppose $A = 30°$, $v_0 = 160$ ft. per sec. Find the horizontal distance from the starting point to the point where the projectile will strike the ground.

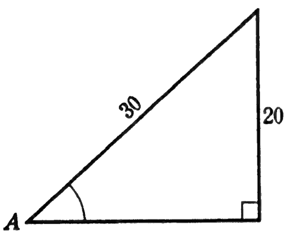


FIG. 173

Clearly we want the value of x for which $y = 0$. Now from (2), $y = 0$ where $t(v_0 \sin A - 16\, t) = 0$ or at $t = 0$ and at $t = (v_0 \sin A)/16$. The first solution $t = 0$ corresponds to the starting point. The second solution is $t = 160(.5)/16 = 5$. Hence the projectile strikes the ground at the end of 5 seconds. From (1) this means that the distance $x = 160 \cdot 5(.866) = 692.8$ ft. approximately.

Remark 1. If more accurate approximations are desired we can use interpolation, or more extensive tables, or both.

Remark 2. How are the tables constructed? A rough table could be made by direct methods. For example, to get cos 40° we might construct an angle of 40° with a protractor as accurately as we could; then we could form a right triangle, measure the adjacent side and hypotenuse as closely as possible and divide one by the other. However, such a method would hardly be satisfactory except for rough work, for we could never be sure as to how accurate our results were; that is, we would never know how many decimal places in our results were really correct. A better method, which is actually used, is based on the following theorem, proved by the calculus. If x is the number of *radians* in the angle, then $\cos x$ is given by the infinite series

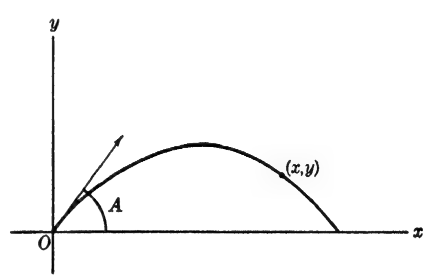


FIG. 174

$$\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{4 \cdot 3 \cdot 2} - \frac{x^6}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2} + \frac{x^8}{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2} - \cdots.$$

Now $40° = .7$ radians approximately. Hence $\cos 40° =$

$$\cos .7 = 1 - \frac{(.7)^2}{2} + \frac{(.7)^4}{4 \cdot 3 \cdot 2} - \frac{(.7)^6}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2} + \cdots.$$

Taking the first three terms of this series we get $\cos 40° = .765$ approximately, which is correct as far as the hundredths place, as we see from the table. Similar infinite series for the other trigonometric functions may be found in text-books on the calculus. By this method we are able to calculate the trigonometric functions as accurately as we wish by taking enough terms of the infinite series. At each stage we can tell how large an error is committed by our approximation. Even with these modern methods, the labor involved in the calculation of tables is great, but the labor saved by having the tables is inestimable.

In the exercises in this chapter, angles should be found correct to the nearest degree and lengths correct to the nearest tenth.

EXERCISES

1. Find the height of a pole whose (horizontal) shadow is 85 ft. long when the angle of elevation of the sun is 51°.

2. Find the length of the shadow of a pole 85 ft. high when the angle of elevation of the sun is 51°.

3. A wire 50 ft. long is stretched from the top of a 35 ft. pole to the ground. What angle does the wire make with the pole?

4. A wire 50 ft. long is stretched from the top of a pole to the ground making an angle of 52° with the pole. How high is the pole?

5. A wire is to be stretched from the top of a 50 ft. pole to the ground making an angle of 48° with the ground. How long should the wire be?

6. A wire is stretched from the top of a pole to a point on the ground 50 ft. from the base of the pole. If the angle between the wire and the ground is 50° what is the height of the pole?

7. From the top of a lighthouse 100 ft. high the angle of depression * (Fig. 175) of a boat B is found to be 23°. How far from the foot F of the lighthouse is the boat?

* The angle between the line of sight and the horizontal.

8. Draw, with a protractor, an angle of 35° on graph paper. By measurement and long division, estimate the values of the trigonometric functions of 35°. Compare with the table.

9. A projectile is fired at an angle of 42° with the horizontal at an initial velocity of 144 ft. per sec. Using the formulas (1) and (2) of example 3 above, find: (*a*) the distance of the point where the projectile strikes the ground from the starting point; (*b*) the maximum height reached by the projectile. (Hint: use the method of section 109 for part (*b*).)

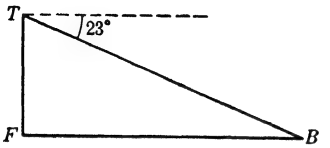


Fig. 175

10. A man drives 500 ft. up a road whose inclination is 30°. How high above his starting level is he?

11. From a point opposite the middle of a battleship known to be 500 ft. long, turned broadside, the angle subtended * by the ship is 6°. How far away is the battleship?

12. If the radius of the earth is 3960 miles, find the radius of the 42nd parallel (of latitude) (Fig. 176).

13. Using the result of exercise 12, find the circumference of the 42nd parallel. How many miles do we cover if we traverse 36° of longitude along the 42nd parallel?

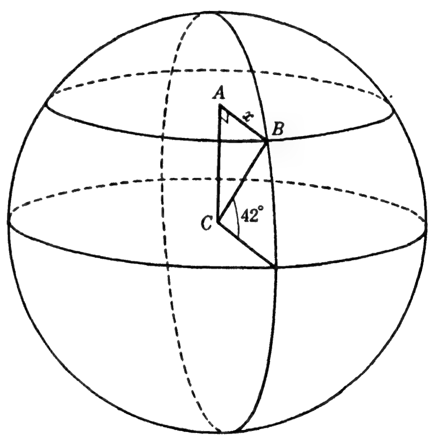


Fig. 176

14. An observer in an airplane 3000 ft. directly above a cannon finds that the angle of depression of an enemy supply base is 32°. Find the horizontal distance between the cannon and the supply base.

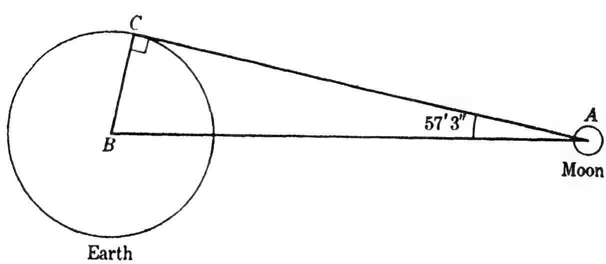


Fig. 177

* The angle **subtended** by an object at a point P is the angle between the lines of sight from P to the extremities of the object, as viewed from P.

15. If the radius of the earth is 3960 miles and angle A (Fig. 177) is found to be $57'3''$, find the distance between the center of the earth and the center of the moon. Suppose that $\sin 57'3'' = .01659$.

16. An airplane is 1 mile above the ocean. How far away from it is the furthest visible point on the ocean (Fig. 178)? Suppose that the radius of the earth is 3960 miles.

17. Find the angle of elevation of the sun if a pole 50 ft. long casts a horizontal shadow 22 ft. long.

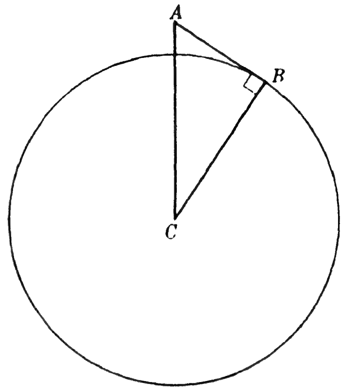


FIG. 178

122. Vectors. Many physical quantities involve not only magnitude but direction as well. For example, to say that a force of 30 lbs. is applied to an object does not specify the force completely; we need to know the direction in which the force is being applied. Velocities are also not completely described unless the direction is given. Such quantities, speaking loosely, are called **vectors.** A vector is conveniently represented by an arrow whose length corresponds to the magnitude and which points in the proper direction. Thus a force of 3 lbs. applied in an easterly direction is represented graphically by an arrow 3 units long pointing east.

It is a well known result of experimental physics that if two forces $\overrightarrow{AC}$ and $\overrightarrow{AB}$ making an angle with each other are applied simultaneously, the resultant force is described with respect to both its magnitude and direction, by the diagonal $\overrightarrow{AD}$ of a parallelogram as in Fig. 179. This is called the *parallelogram law for the composition of forces.* Velocities also may be represented as vectors as may many other physical quantities. Vectors, like other parts of pure mathematics, may be applied to (and hence serve to unify) many different concrete physical interpretations.

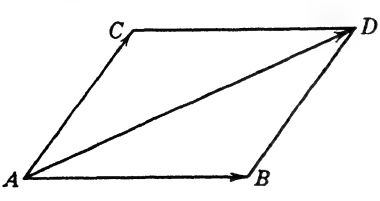


FIG. 179

The two given vectors are called **components,** and the diagonal vector is called the **resultant.** We shall confine ourselves in this section to components which are at right angles to each other.

Example 1. A body is acted upon simultaneously by a force of 50 lbs. due north, and a force of 120 lbs. due east. Find the magnitude and direction of the resultant force. See Fig. 180.

By the Pythagorean theorem, $(AD)^2 = 50^2 + 120^2$; hence $AD = 130$. Clearly $\tan x = 50/120 = .4167$. From the table of tangents we find $x = 23°$ approximately. Hence the resultant force is one of 130 lbs. acting in a direction 23° north of east.

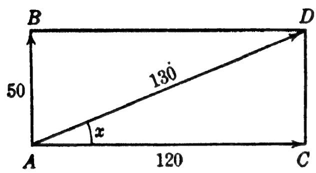


Fig. 180

Example 2. A boat pointed east travels at the rate of 24 miles per hour under its own power and a wind blows it north at the rate of 10 miles per hour. Find the magnitude and direction of the velocity of the boat. See Fig. 181.

The resultant of the two given components is a velocity of 26 miles per hour, and $\cot x = 2.4$. From the table we find $x = 23°$ approximately. Hence the boat moves in a direction 23° north of east, with a speed of 26 miles per hour.

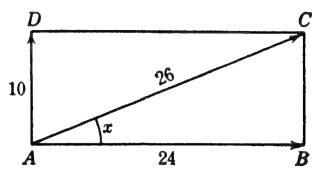


Fig. 181

Example 3. A body weighing 100 lbs. is on a plane inclined at an angle of 30° to the horizontal. What force must be applied to the body to keep it from sliding down the plane? Friction is to be neglected. See Fig. 182.

The weight of the body means that the force of gravity acting on it is 100 lbs. directed downward. We regard $\overrightarrow{BG}$ as the resultant of the two perpendicular forces $\overrightarrow{BD}$ and $\overrightarrow{BR}$. Clearly $BD/100 = \cos 60°$ or $BD = 50$. Hence the force pushing the body downhill is 50 lbs. in the direction $\overrightarrow{BD}$. Thus we must apply a force of 50 lbs. in the opposite direction $\overrightarrow{DB}$.

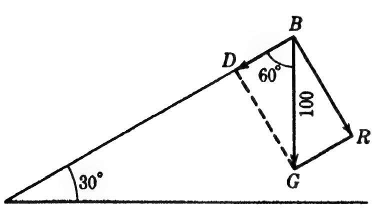


Fig. 182

EXERCISES

In all the following exercises, friction is to be neglected.

1. A body is acted upon by a force of 75 lbs., due east, and a force of 100 lbs., due north. Find the magnitude and direction of the resultant force.

2. A river flows at the rate of 3 miles per hour. A man who rows at the rate of 4 miles per hour in still water sets out directly across the stream. Find the magnitude and direction of his actual velocity.

3. A block weighing 200 lbs. is on a plane inclined at 45° to the horizontal. Find the magnitude and direction of the force needed to keep it from sliding downhill.

4. An airplane pointed due east is flying at the rate of 150 miles per hour. A wind blowing due south, blows it south at the rate of 30 miles per hour. Find the magnitude and direction of its actual velocity.

5. An airplane takes off at an angle of 5° while moving at the rate of 80 miles per hour. How fast is it rising? How fast is it moving forward (that is, horizontally)?

6. A man walks north across a railway car at the rate of 2 miles per hour while the car proceeds east at the rate of 10 miles per hour. What is the direction and magnitude of his motion with respect to the surface of the earth?

7. What force is required to drag a 2000 lb. automobile up a ramp inclined 25° from the horizontal?

8. What is the largest weight which a man can keep from sliding down a ramp inclined 30° from the horizontal if he can exert a pull of 150 lbs.?

9. A weight of 200 lbs. is to be kept from sliding down a ramp by a man whose maximum pull is 125 lbs. What is the largest possible angle at which the ramp may be inclined to the horizontal?

123. Directed angular measure. One of the inconvenient features of our angular measures is that we cannot freely add or subtract them. For example, we cannot say that an angle whose measure is 100° plus an angle whose measure is 110° gives an angle whose measure is 210°, since no angular measure can be greater than 180° by our definition. Similarly we cannot subtract an angular measure of 110° from an angular measure of 100° to get one of − 10° since "negative" angular measures have not been defined. However, for many purposes it is convenient to extend the notion of angular measure to allow for *directed* angular measures of any positive or negative number of degrees like − 960° or + 750°, just as it is convenient to use directed numbers for coordinates. To do this we agree that counter-clockwise rotation shall be regarded as positive and clockwise as negative. If we imagine the rotation beginning at one ray of an angle and ending at the other we shall

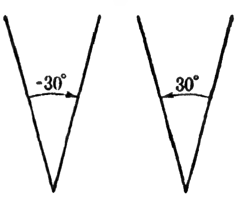


FIG. 183

call the first ray the **initial ray** and the other the **terminal ray.** Having decided which of the two rays of an angle is to be its initial ray, we shall mean by a **directed angular measure** of the given angle, with the given initial ray, the (directed) number of degrees in any rotation starting at the initial ray and ending at the terminal ray of the angle. The directed angular measure of an angle is still a function of the angle, with specified initial ray, but it is no longer a single-valued function, since the same angle may have different directed angular measures. For example, an angle, with specified initial ray, having a directed angular measure of 30°, also has the directed angular measures 390°, 750°, − 330°, − 690°, and so on, since rotating through 360° any number of times in either direction yields the same terminal ray. In fact, *if an angle (with specified initial ray) has a directed angular measure of x degrees, it also has all the directed angular measures given by $x + n \cdot 360$ degrees where n is any integer, positive, negative, or zero.* It is customary to speak of an *angle* of 390° or − 330° instead of a directed angular measure of 390° or − 330°. In the figure a curved arrow is used to indicate the amount and direction of the rotation.

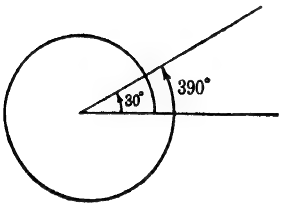


Fig. 184

Since all these directed angular measures correspond to the same angle, you may wonder why we bother to distinguish among them. But if you recall that our directed angular measure corresponds to the intuitive notion of rotation, you will see that there is a very practical distinction between + 750° and − 690°; if your car is parked near the edge of a cliff, you will be seriously concerned with the question of which of these rotations your wheels will make.

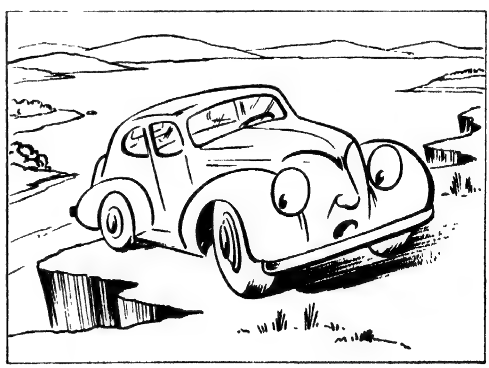


Fig. 185

EXERCISES

Name 5 other directed angular measures, three positive and two negative, possessed by an angle with the directed angular measure:

1. $+10°$. **2.** $-10°$. **3.** $60°$. **4.** $150°$. **5.** $210°$. **6.** $300°$.

124. Trigonometric functions of any angle. Our previous definitions of the trigonometric functions in terms of the sides of a right triangle are clearly inapplicable to any angle other than an acute one; for an angle of 120°, say, cannot be put inside a right triangle. We shall now define the trigonometric functions of any angle (that is, any directed angular measure).

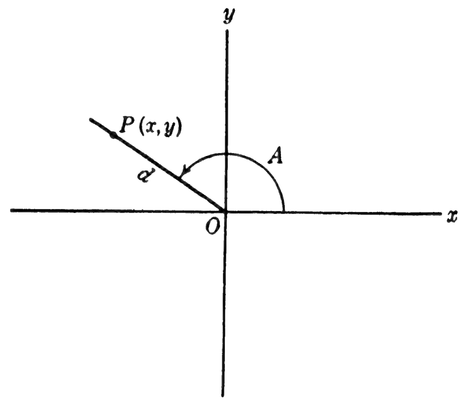


FIG. 186

DEFINITION. *Suppose any angle (with specified initial ray) has been given. Introduce an ordinary rectangular coordinate system into the plane with the origin at the vertex of angle A and the initial ray of angle A as the positive ray of the x-axis. Choose any point P whatever on the terminal ray of angle A, and let (x,y) be the coordinates of P and d its distance from the origin* (Fig. 186). *Then we define*

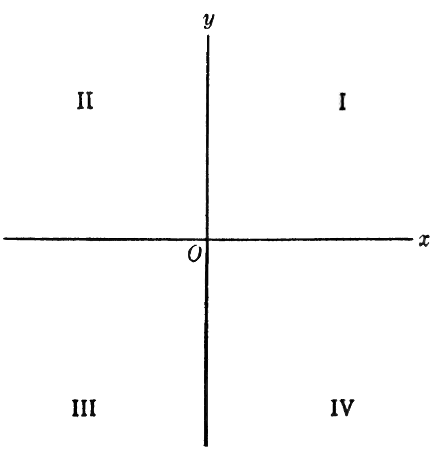


FIG. 187

$$\begin{aligned}
\sin A &= y/d = \text{ordinate/distance} \\
\cos A &= x/d = \text{abscissa/distance} \\
\tan A &= y/x = \text{ordinate/abscissa} \\
\csc A &= d/y = \text{distance/ordinate} \\
\sec A &= d/x = \text{distance/abscissa} \\
\cot A &= x/y = \text{abscissa/ordinate.}
\end{aligned}$$

These quantities are really functions of the angle A alone. In particular, they do not depend on our choice of the point P; for if we chose a different point $P'(x',y')$ on the terminal ray of

A at a distance d' from the origin, it is clear that x' and y' would have the same signs as x and y respectively and, because of similar triangles, the ratios y'/d', x'/d', y'/x', etc., would be equal respectively to the ratios y/d, x/d, y/x, etc. Distance is always positive.

Clearly *any trigonometric function of an angle of $A°$ is equal to the same trigonometric function of any angle of $A + n \cdot 360$ degrees where n is any integer,* since all these angles have the same terminal ray. For example, $\sin 30° = \sin 390° = \sin(-330°)$.

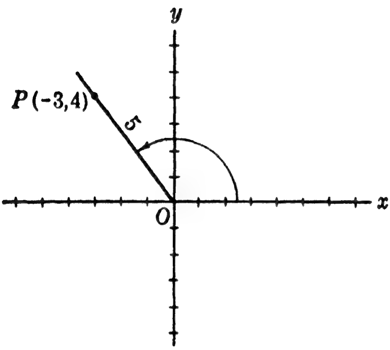


FIG. 188

After the coordinate system has been introduced, the plane is divided into four **quadrants,** which are always numbered as in Fig. 187. An angle is said to be "in" the quadrant in which its terminal ray lies. For positive acute angles, our new definitions yield the same results as those of section 120, since x and y are both positive for any point in the first quadrant, and consequently are nothing more than the lengths of the sides adjacent and opposite to A, respectively. Therefore, no conflict arises between our old and new definitions. Our new definitions may be called "generalizations" of the old ones since they agree with the old ones for all angles to which the old ones may be applied (that is, acute angles) but are applicable to a wider class of angles.

Example 1. Find the cosine of the angle whose terminal ray passes through the point $P(-3,4)$. See Fig. 188.

The distance $d = PO$ is clearly 5 since, by the Pythagorean theorem, $d^2 = x^2 + y^2 = (-3)^2 + 4^2 = 25$. Thus $\cos A = -3/5$.

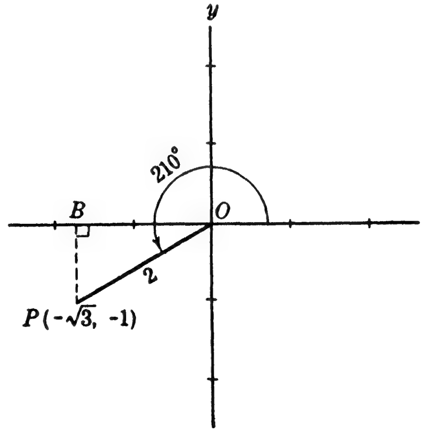


FIG. 189

Example 2. Find the trigonometric functions of 210°. Clearly if we choose a point P on the terminal ray whose distance from O is 2 units, the coordinates of P are $(-\sqrt{3}, -1)$ since BPO is a 30°–60°–90° triangle. Hence (Fig. 189)

$$\begin{aligned} \sin 210^\circ &= -1/2 & \csc 210^\circ &= -2 \\ \cos 210^\circ &= -\sqrt{3}/2 & \sec 210^\circ &= -2/\sqrt{3} \\ \tan 210^\circ &= 1/\sqrt{3} & \cot 210^\circ &= \sqrt{3}. \end{aligned}$$

Example 3. Find the trigonometric functions of 90°. Choose a point P on the terminal ray with coordinates (0,1). The distance d is then 1. Hence (Fig. 190)

$\sin 90^\circ = 1$ $\qquad$ $\csc 90^\circ = 1$
$\cos 90^\circ = 0$ $\qquad$ $\sec 90^\circ$ does not exist
$\tan 90^\circ$ does not exist $\qquad$ $\cot 90^\circ = 0.$

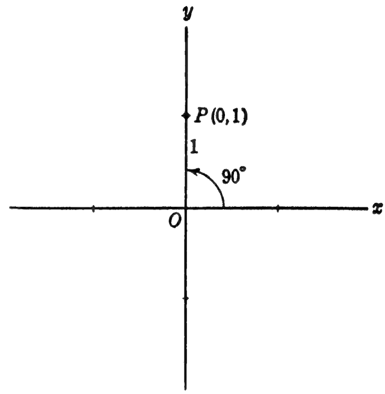


FIG. 190

EXERCISES

Find the trigonometric functions of an angle whose terminal ray contains the point:

1. (3,4). **2.** (– 3,4). **3.** (– 3,– 4). **4.** (3,– 4). **5.** (4,– 3).
6. (0,– 1). **7.** (1,0). **8.** (– 1,0). **9.** (5,12). **10.** (– 12,5).

Find the trigonometric functions of the following angles:

11. 30°. **12.** 150°. **13.** 240°. **14.** 330°. **15.** 510°.
16. – 60°. **17.** – 120°. **18.** – 240°. **19.** – 300°. **20.** – 420°.
21. 45°. **22.** 135°. **23.** 225°. **24.** 315°. **25.** 765°.
26. 180°. **27.** 270°. **28.** 360°. **29.** 0°.

In which of the quadrants may angle A terminate if:

30. $\sin A > 0$. **31.** $\cos A < 0$. **32.** $\tan A > 0$.
33. $\csc A < 0$. **34.** $\sec A > 0$. **35.** $\cot A < 0$.

Find all the trigonometric functions of angle A if:

36. $\sin A = 3/4$ and the terminal ray of A is in quadrant I.
37. $\sin A = 3/4$ and the terminal ray of A is not in quadrant I.
38. $\cos A = 5/12$ and the terminal ray of A is not in quadrant I.
39. $\tan A = 3/5$ and the terminal ray of A is not in quadrant I.
40. $\tan A = -3/5$ and the terminal ray of A is not in quadrant II.

125. The line-values and the names of the trigonometric functions. In the preceding section we agreed to *call* the ratio y/d by the name "sine of A." Needless to say, if we had wished to, we could have called it the "abracadabra of A." A student who is tempted to ask "How do we know that y/d *is* the sine of

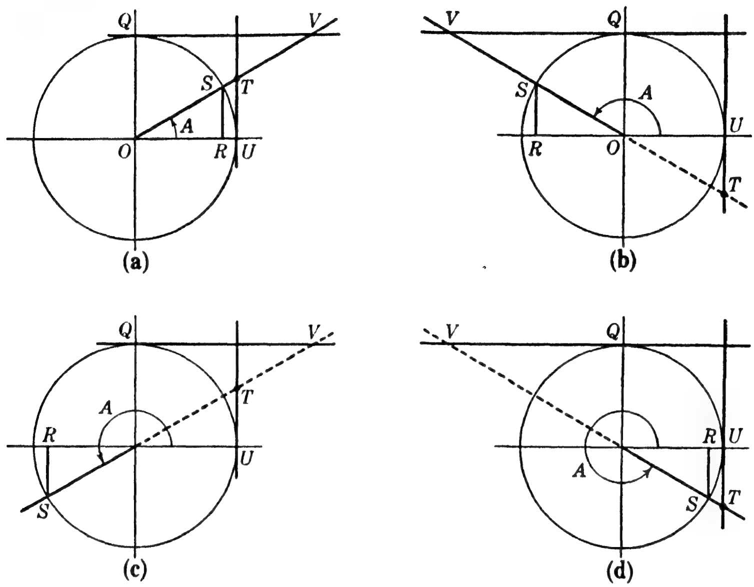


FIG. 191

A?" is in the same class as the Gushing Lady who came up to a well-known astronomer after a public lecture, as gushing ladies do, to tell him how clear his talk was and how much she enjoyed it. She concluded by saying, "I understood how you found out the distances of the stars and all that, but there was one thing I didn't grasp. How did you ever find out their names?"

The choice of the names of the trigonometric functions is, logically, an arbitrary matter. But there does remain the historical or psychological question, "Why did we happen to choose these particular names rather than others?" Some light will be thrown on this question by examining the following geometric interpretation of the trigonometric functions.

Consider a circle of radius one unit with center at the origin, and let the positive ray of the x-axis lie along the initial ray of angle A, as in the preceding section. Then (Fig. 191) draw tangents to the circle at Q and U. Let S be the point where the

terminal ray of angle A meets the circle. Drop a perpendicular from S to the x-axis meeting the x-axis at R. Produce the line OS until it meets the horizontal tangent line at V and the vertical tangent line at T. Let us agree that horizontal line-segments are to be counted positively or negatively according as they extend to the right or left of the y-axis; and that vertical line segments are to be counted positively or negatively according as they extend up or down from the x-axis. Segments which have been obtained by producing the terminal ray backwards through O will be counted negatively. They are dotted in Fig. 191. Now let us consider the trigonometric functions of A.

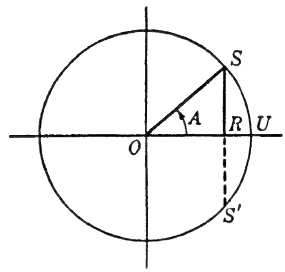


FIG. 192

Clearly $OU = OQ = OS = 1$. Taking account of our agreements as to signs, $\tan A = UT/OU = UT$. That is, the tangent of A is represented by the segment cut off by the terminal ray of A (produced if necessary) on the vertical tangent line.

In the same way, we have $\cot A = QV/OQ = QV$. That is, the cotangent of A is represented by the segment cut off by the terminal ray of A (produced if necessary) on the horizontal tangent line. It is, of course, the tangent of the complement of A.

Similarly $\sec A = OT/OU = OT$. The name secant may have been suggested by the fact that the segment OT "cuts" the circle; the word "secant" comes from a Latin word which means "cutting."

Likewise, $\csc A = OV/OQ = OV$. This is, of course, the secant of the complement of A.

The names of the four functions considered so far have had fairly plausible geometric motivation. The case of sine and cosine is not so easily explained. If we can explain the origin of the name "sine," then "cosine" is simply the sine of the complement. Clearly, on Fig. 191, we have $\sin A = RS/OS = RS$ and $\cos A = OR/OS = OR$. But this scarcely explains the word sine, which comes from the Latin "sinus," meaning fold, cavity, or bay. The fact is that the word sinus is due to a mistranslation, although historians are not agreed as to just what mistranslation was committed. Note (Fig. 192) that the sine of A is RS which

is half the chord SS'. The ancients, however, were accustomed to use the entire chord SS' as the sine instead of the half-chord RS. Now, one plausible explanation of the origin of the word sinus is that the Hindu word for sine meant "bowstring," an idea clearly suggested by the figure of the chord SS' together with the arc SUS'; but because of similarity in sound, Arab translators confused their word for bowstring with a word meaning fold or bay, from which Latin translators got the word sinus.

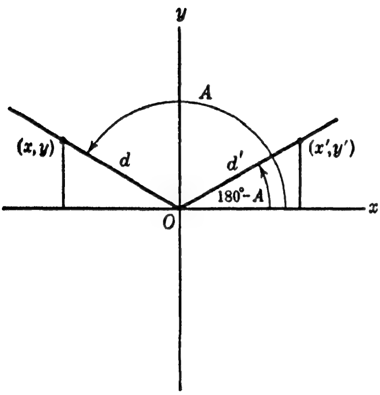


FIG. 193

126. Reduction formulas. Since the table gives the values of the trigonometric functions for acute angles only, how can we look up the sine of 160°, say? Clearly we have somehow to express the trigonometric functions of any angle in terms of the trigonometric functions of an acute angle (that is, an angle between 0° and 90°).

Case I. If A terminates in quadrant I, there is an angular measure between 0° and 90° which can be looked up having the same terminal ray and therefore the same trigonometric functions. For example, $\sin 380° = \sin 20°$ which can be looked up. Hence $\sin 380° = .3420$.

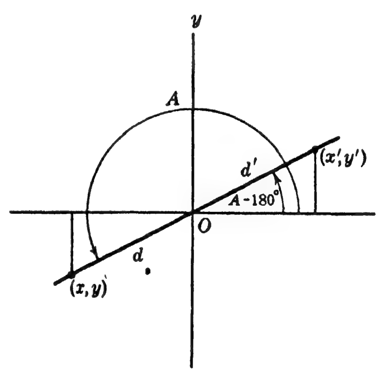


FIG. 194

Case II. If A terminates in quadrant II, $180° - A$ terminates in quadrant I. For example, if $A = 160°$, $180° - A = 20°$. Choose a point (x,y) on the terminal ray of A at a distance d from the origin. Choose a point (x',y') on the terminal ray of $180° - A$ such that its distance from the origin is $d' = d$. Then clearly (Fig. 193), $y = y'$ and $x = -x'$. Hence,

$$\begin{aligned}
\sin A &= y/d = y'/d' = \sin(180° - A)\\
\cos A &= x/d = -x'/d' = -\cos(180° - A)\\
\tan A &= y/x = -y'/x' = -\tan(180° - A)\\
\csc A &= d/y = d'/y' = \csc(180° - A)\\
\sec A &= d/x = -d'/x' = -\sec(180° - A)\\
\cot A &= x/y = -x'/y' = -\cot(180° - A).
\end{aligned}$$

For example, $\sin 160° = \sin 20°$, $\cos 160° = -\cos 20°$, etc., and the functions of 20° can be looked up in the table. Hence $\sin 160° = .3420$, $\cos 160° = -.9397$, $\tan 160° = -.3640$, $\csc 160° = 2.924$, etc.

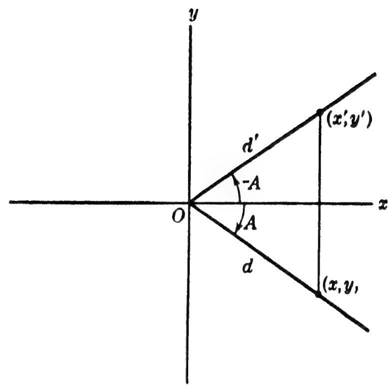


FIG. 195

Case III. If A terminates in quadrant III, then $A - 180°$ terminates in quadrant I. Choose points (x,y) and (x',y') on the terminal ray of A and $A - 180°$, respectively, such that $d = d'$ (Fig. 194). Then $x = -x'$, $y = -y'$. Hence

$$\begin{aligned}
\sin A &= y/d = -y'/d' = -\sin(A - 180°)\\
\cos A &= x/d = -x'/d' = -\cos(A - 180°)\\
\tan A &= y/x = -y'/-x' = y'/x' = \tan(A - 180°)\\
\csc A &= d/y = -d'/y' = -\csc(A - 180°)\\
\sec A &= d/x = -d'/x' = -\sec(A - 180°)\\
\cot A &= x/y = -x'/-y' = x'/y' = \cot(A - 180°).
\end{aligned}$$

For example, $\sin 200° = -\sin 20°$, $\tan 200° = \tan 20°$. Hence, $\sin 200° = -.3420$, $\cos 200° = -.9397$, $\tan 200° = .3640$, $\csc 200° = -2.924$, etc.

Case IV. If A terminates in quadrant IV then $-A$ terminates in quadrant I. Choose points (x,y) and (x',y') on the terminal rays of A and $-A$ respectively such that $d = d'$ (Fig. 195). Then $x = x'$, $y = -y'$. Hence

$$\begin{aligned}
\sin A &= y/d = -\, y'/d' = -\sin(-A)\\
\cos A &= x/d = x'/d' = \cos(-A)\\
\tan A &= y/x = -\, y'/x' = -\tan(-A)\\
\csc A &= d/y = -\, d'/y' = -\csc(-A)\\
\sec A &= d/x = d'/x' = \sec(-A)\\
\cot A &= x/y = -\, x'/y' = -\cot(-A).
\end{aligned}$$

For example, if $A = -20°$ then $-A = 20°$. Thus $\sin(-20°) = -\sin 20°$, $\cos(-20°) = \cos 20°$. Hence, $\sin(-20°) = -.3420$, $\cos(-20°) = .9397$.

The formulas obtained in this section are called **reduction formulas** because, by means of them, the problem of finding the trigonometric functions of any angle is reduced to the problem of finding the trigonometric functions of an angle between 0° and 90°, which may be looked up in the table, and prefixing the proper sign. That is, they enable us to get along with a table giving the values of the trigonometric functions for acute angles only.

EXERCISES

Find, using the reduction formulas and the tables, the values of the trigonometric functions of:

1. 170°. **2.** 190°. **3.** $-10°$. **4.** 350°. **5.** 530°.
6. $-730°$. **7.** 375°. **8.** 155°. **9.** 217°. **10.** 310°.
11. $-165°$ **12.** $-305°$.

13. Make a table giving the values of the trigonometric functions for 0°, 15°, 30°, 45°, 60°, $\cdots$, 360° (all angles 15° apart up to 360°).

Find angle A if:

14. $\sin A = .3420$, A not in quadrant I.
15. $\cos A = -.8387$, A between 0° and 180°.
16. $\sin A = 1/2$, A not in quadrant I.
17. $\cos A = -1/2$, A between 0° and 180°.
18. $\sin A = .6018$, A not in quadrant I.
19. $\cos A = -.8391$, A between 0° and 180°.
20. $\sin A = \sqrt{3}/2$, A not in quadrant I.
21. $\cos A = -\sqrt{3}/2$, A between 0° and 180°.
22. $\cos A = -\sqrt{2}/2$, A between 0° and 180°.
23. $\cos A = -.3420$, A between 0° and 180°.
24. $\tan A = 1$, A not in quadrant I.

127. The solution of oblique triangles. To find the height of a pole BC (Fig. 196), we have only to measure off any convenient distance AC from the foot of the pole, measure the angle A and write $BC/AC = \tan A$. For example, if $AC = 100$ ft. and $A = 38°$ then $BC = AC \tan A = 100(.7813) = 78.13$ ft. But to find the height h of a mountain we cannot proceed in the same way. To apply this method we should have to measure the distance AC (Fig. 197). But any attempt to measure directly the distance AC will obviously be sabotaged by the mountain. To circumvent the mountain's irritating obstructionist tactics, we proceed as follows. Measure angle A; suppose it is 30°. Move from A toward the mountain any convenient distance, say 1000 ft.,* to D (Fig. 198). Measure angle BDC; suppose it is 70°. Clearly, if we only knew the distance $a = BD$ we would be able to find h from the relation $h/a = \sin 70°$. In the next section we shall develop a way of finding a from what we already know about the oblique † triangle ABD.

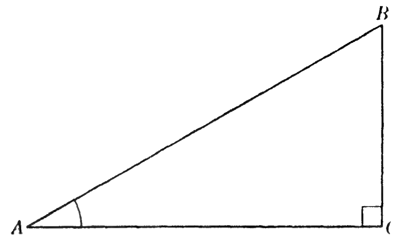


FIG. 196

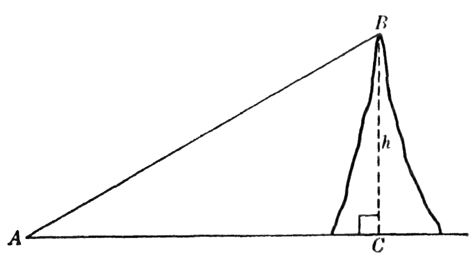


FIG. 197

Note that in triangle ABD we know two angles and the included side, since $\measuredangle A = 30°$, $AD = 1000$ ft., and $\measuredangle ADB$ is clearly 110° since it is supplementary to $\measuredangle BDC$. Now this is enough to determine or fix or specify the size and shape of the triangle completely, since, if any other triangle has two angles and the included side equal to 30°, 110°, and 1000 ft. respectively, the second triangle must be congruent to the given triangle ABD. Since the size and shape of the triangle are determined by the data, we may reasonably expect to find the remaining sides and angles of triangle ABD.

* The plain from A to D is assumed to be horizontal, as is usually the case—in text books.

† An **oblique triangle** is one which is not a right triangle.

In general, whenever we know enough about the sides and angles of a triangle to guarantee that any other triangle having the given measurements must be congruent to it, we shall expect to be able to compute the remaining sides and angles. From elementary geometry, we recall the following theorems about the congruence of triangles.

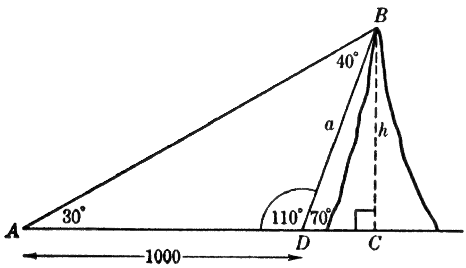


FIG. 198

Two triangles are congruent if:

(1) two angles and the included side of one are equal respectively to two angles and the included side of the other;

(2) two sides and the included angle of one are equal respectively to two sides and the included angle of the other;

(3) three sides of one are equal respectively to three sides of the other.

Note that (1) can be modified to read "any two angles and one side" since if two angles of one triangle are equal to two angles of another triangle, then the third angles are equal. This is so because the sum of the angles of any triangle is 180°.

Hence we shall expect to be able to calculate the remaining sides and angles of any triangle when we are given:

Case I. any two angles and a side;

Case II. two sides and the included angle;

Case III. three sides.

To **solve** a triangle means to find the remaining sides and angles when some of them are known. In the next section we shall dispose of Case I and in the following section Cases II and III.

Exercise. The height h of the mountain in our illustrative example can be found by using two equations and two unknowns, without the easier method of the next section. Do it. (Hint: h is one unknown; let $k = DC$ be the other. Use triangles BCD and ABC, Fig. 198, to get two equations.)

128. The law of sines. Consider any oblique triangle ABC. From B draw a perpendicular BD to AC. Either BD falls within

the triangle ABC or not (Fig. 199).* We shall treat each case separately.

Case (a). From right triangle ABD (Fig. 199a) we obtain $h/c = \sin A$ or

(1) $$h = c \sin A.$$

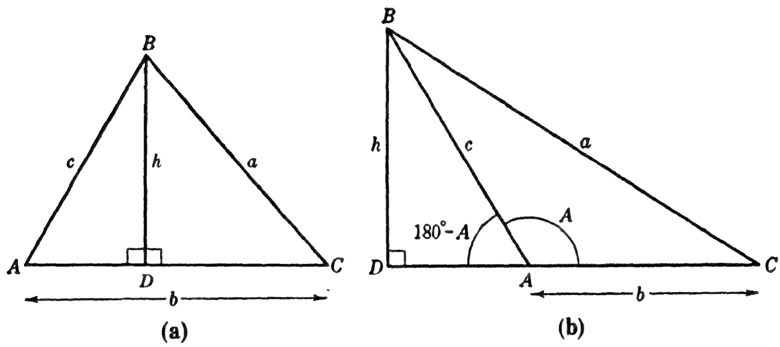


FIG. 199

From right triangle BDC we obtain $h/a = \sin C$ or

(2) $$h = a \sin C.$$

From (1) and (2) we have

$$a \sin C = c \sin A.$$

Dividing both sides by $\sin A \sin C$, we obtain

(3) $$\frac{a}{\sin A} = \frac{c}{\sin C}.$$

Case (b). From right triangle BDC (Fig. 199b) we obtain $h/a = \sin C$ or $h = a \sin C$ as before. From right triangle BDA we have $h/c = \sin (180° - A)$. But $\sin (180° - A) = \sin A$. Hence $h/c = \sin A$ or $h = c \sin A$ as before. As in Case (a) we obtain equation (3).

By dropping perpendiculars from A to BC instead of from B to AC we would obtain in the same fashion the equation

(4) $$\frac{b}{\sin B} = \frac{c}{\sin C}.$$

From (3) and (4) we have the following theorem, known as the **law of sines.**

* The perpendicular BD cannot fall exactly along one side of the triangle ABC since ABC is oblique by hypothesis.

THEOREM. *In any triangle,*

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C}.$$

Or, any side divided by the sine of the opposite angle is equal to any other side divided by the sine of **its** *opposite angle.*

Remark. We have proved the theorem for oblique triangles. The reader can verify at once that it remains true even if one of the angles is 90°, recalling the fact that $\sin 90° = 1$.

Exercise. Complete the proof in all its details.

By means of the law of sines we can solve a triangle given two angles and one side.

Example 1. Given triangle ABC with $A = 40°$, $B = 80°$, and $a = 10$ ft. Then $C = 60°$. To find b we write $b/\sin B = a/\sin A$ or

$$\frac{b}{\sin 80°} = \frac{10}{\sin 40°}$$

or $$b = \frac{10 \sin 80°}{\sin 40°} = \frac{10(.9848)}{.6428} = 15.3 \text{ ft. approx.}$$

To find c we would use similarly the relation $c/\sin C = a/\sin A$.

Example 2. Let us now complete the problem of finding the height of the mountain begun in section 127 (Fig. 198). Clearly angle $ABD = 40°$. By the law of sines we have, from triangle ABD,

$$\frac{a}{\sin 30°} = \frac{1000}{\sin 40°}.$$

Hence $$a = \frac{1000 \sin 30°}{\sin 40°} = \frac{1000(.5000)}{.6428}$$
$$= 777.8 \text{ ft. approx.}$$

Having found a we obtain from the right triangle BCD (Fig. 198) the relation $h/777.8 = \sin 70°$ or

$$h = 777.8 \sin 70° = 777.8(.9397)$$
$$= 731 \text{ ft. approx.}$$

EXERCISES

Solve the triangle ABC, given that:

1. $A = 42°$, $B = 73°$, $c = 15$ ft.

2. $B = 53°$, $C = 62°$, $c = 32$ ft.

3. $A = 103°$, $C = 48°$, $a = 25$ ft.

4. $A = 26°$, $B = 132°$, $c = 20$ ft.

5. To measure the length AB of a projected bridge across a canyon, a distance of 100 yards is measured from A to a point C and the angles BAC and ACB are found to be 82° and 43° respectively. Find AB.

6. To measure the distance from a cannon C to an enemy fortification F, a distance of 1000 yards is measured from C to a point P, and the angles FCP and CPF are found to be 77° and 62°, respectively. Find CF.

7. From two observation points A and B, 5000 ft. apart, on a straight road an airplane C is seen directly over the road. If, at a given instant the angle of elevation of the airplane at A is 73°, and the angle of elevation of the airplane at B is 67°, find: (*a*) the distance of the airplane from A; (*b*) the distance of the airplane from B; (*c*) the airplane's altitude.

8. A ship S is observed from two points A and B on a straight beach, 1000 ft. apart. If $\measuredangle SAB = 82°$ and $\measuredangle SBA = 73°$, find: (*a*) the distance SA; (*b*) the distance SB; (*c*) the distance of the ship from the shore.

9. To measure the height of a mountain, its angle of elevation from a point A on a level plain is found to be 34°. Proceeding toward the mountain, a distance $AB = 1000$ ft. is measured off. At B the angle of elevation of the mountain is found to be 58°. Find the height of the mountain.

10. Outline a method for finding the height of a tower located on the opposite bank of a river, without crossing the river. Invent your own numbers and solve.

129. The law of cosines. Consider the oblique triangle ABC. Draw the line BD perpendicular to AC. Either BD falls within the triangle ABC or not (Fig. 200). (Since ABC is oblique BD

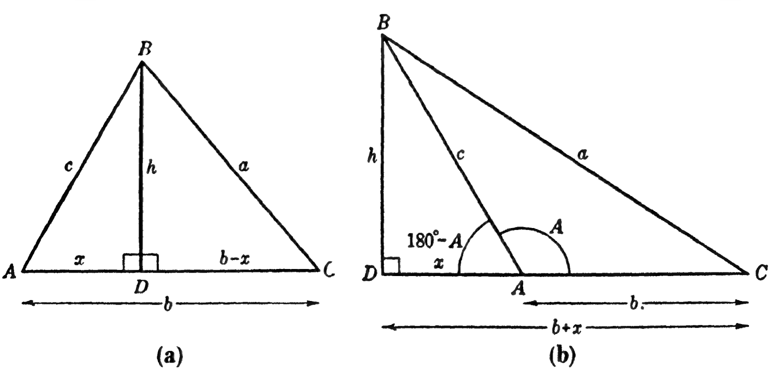


FIG. 200

cannot fall along one side of triangle ABC.) We shall treat each case separately.

Case (a). From right triangle ABD (Fig. 200a) we have $c^2 = x^2 + h^2$ or

(1) $$h^2 = c^2 - x^2.$$

From right triangle BDC we have $a^2 = (b - x)^2 + h^2$ or

(2) $$h^2 = a^2 - (b - x)^2.$$

From (1) and (2) we obtain

$$a^2 - (b - x)^2 = c^2 - x^2,$$

or

$$a^2 - b^2 + 2\,bx - x^2 = c^2 - x^2,$$

or

(3) $$a^2 = b^2 + c^2 - 2\,bx.$$

From right triangle ABD, we have $x/c = \cos A$, or

(4) $$x = c \cos A.$$

Substituting (4) in (3), we have

(5) $$a^2 = b^2 + c^2 - 2\,bc \cos A.$$

Case (b). From right triangle BDA (Fig. 200b), we have $c^2 = h^2 + x^2$, or

(6) $$h^2 = c^2 - x^2.$$

From right triangle BDC we have $a^2 = h^2 + (b + x)^2$ or

(7) $$h^2 = a^2 - (b + x)^2.$$

Hence,

$$a^2 - (b + x)^2 = c^2 - x^2,$$

or

$$a^2 - b^2 - 2\,bx - x^2 = c^2 - x^2,$$

or

(8) $$a^2 = b^2 + c^2 + 2\,bx.$$

From right triangle BDA, we have $x/c = \cos (180° - A)$ or $x = c \cos (180° - A)$. But $\cos (180° - A) = -\cos A$. Hence,

(9) $$x = -c \cos A.$$

Substituting (9) in (8) we obtain (5) as before.

By dropping perpendiculars from A and C, respectively, we

would get the remaining formulas in the following theorem, known as the **law of cosines.**

THEOREM. *In any triangle ABC, we have*

$$a^2 = b^2 + c^2 - 2\,bc\cos A,$$
$$b^2 = a^2 + c^2 - 2\,ac\cos B,$$
$$c^2 = a^2 + b^2 - 2\,ab\cos C.$$

Or, the square of any side is equal to the sum of the squares of the other two sides minus twice their product times the cosine of the angle included between them.

Remark. We have proved the theorem for oblique triangles. But it is true for right triangles as well. For example, if in the last equation angle C were 90° we would have $c^2 = a^2 + b^2 - 2\,ab \cos 90°$. But $\cos 90° = 0$. Hence, $c^2 = a^2 + b^2$, which is nothing but the Pythagorean theorem. Therefore, the law of cosines may be said to be a generalization of the Pythagorean theorem, since it includes the Pythagorean theorem as a special case.

We may use the law of cosines to solve a triangle given three sides or given two sides and the included angle, as follows.

Example 1. Given triangle ABC with $a = 3$, $b = 5$, $c = 7$. Let us find angle C. We have

$$c^2 = a^2 + b^2 - 2\,ab\cos C.$$

Hence $$7^2 = 3^2 + 5^2 - 2 \cdot 3 \cdot 5 \cos C,$$

or $$49 = 34 - 30 \cos C,$$

or $$\cos C = -1/2 = -.5000.$$

Hence, $$C = 120°.$$

Using the other formulas in our theorem similarly we would find angles B and A.

Example 2. Given triangle ABC with $c = 5$, $b = 3$, and $A = 120°$. Let us find a. We have,

$$a^2 = b^2 + c^2 - 2\,bc\cos A,$$

or $$a^2 = 3^2 + 5^2 - 2 \cdot 3 \cdot 5 \cos 120°$$
$$= 34 - 30(-1/2) = 49.$$

Therefore, $a = 7$. Now that we know all three sides the remaining angles may be found by the method of example 1.

EXERCISES

Solve the triangle ABC, given that:

1. $a = 20, \quad b = 12, \quad c = 28.$
2. $a = 35, \quad b = 27, \quad c = 24.$
3. $a = 15, \quad b = 9, \quad C = 120°.$
4. $b = 20, \quad c = 15, \quad A = 67°.$
5. $a = 42, \quad c = 37, \quad B = 79°.$
6. $a = 15, \quad b = 26, \quad C = 112°.$
7. $b = 12, \quad c = 17, \quad A = 123°.$
8. $a = 38, \quad c = 25, \quad B = 105°.$

9. A triangular plot of ground is 215 ft., 185 ft., and 125 ft., respectively, on its three sides. Find the three angles included between its sides. Find its area.

10. To find the length AB of a proposed tunnel through a mountain from point A to point B, a point C is found from which A and B are both visible. If $AC =$ 500 ft., $BC = 350$ ft. and $\measuredangle ACB = 73°$, find AB.

11. To find the distance CT from a cannon C to a target T, invisible from C, an observer is stationed at a place P visible from both C and T. If $TP = 2000$ yds., $CP = 1500$ yds., and $\measuredangle CPT = 84°$, find CT.

12. A force of 12 lbs. and a force of 17 lbs. act on a body simultaneously. If the angle between their directions is 30°, find: (*a*) the magnitude of the resultant force; (*b*) the angle the resultant makes with each component.

13. Two forces of 30 lbs. and 50 lbs., respectively, have an included angle of 60°. Find the magnitude of the resultant force, and the angle it makes with each component.

14. A boat moves across a river in a direction 40° north of east with a speed of 10 miles per hour. The current of the river runs due East with a speed of 5 miles per hour. Find the speed of the boat in still water.

15. The distance from a boat B to two points A and C on the shore are known to be 500 yds. and 400 yds., respectively. If angle ABC is 53°, find the distance AC.

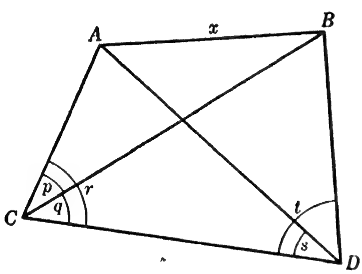


FIG. 201

16. To find the distance between two inaccessible points A and B, we select two observation points C and D, 100 yds. apart, (Fig. 201) and measure angles $p, q, r, s, t.$* Suppose $p = 33°$, $q = 42°$, $r = 75°$, $s = 37°$, $t = 78°$. Find AB. (Hint: From triangle BCD we can find BC. From triangle ACD we can find AC. Then from triangle ABC we can find AB.)

130. The graphs of the trigonometric functions. Periodicity. With the help of section 126, we can draw the graph of the function $y = \sin x$, say, where x is the number of degrees. From the

* If A, B, C, D are in the same plane $r = p + q$. But our method is applicable even if they are not in the same plane.

table, we can plot the point (x,y) where $y = \sin x$ for every angle 0°, 15°, 45°, 60°, 75°, 90°, (Fig. 202). By means of the reduction formulas (see exercise 13, section 126) we can plot the points for all angles 15° apart from 90° to 360° in Fig. 203. Since $\sin x = \sin (x + n \cdot 360°)$, where n is any integer, the curve merely repeats itself in every interval of 360°.

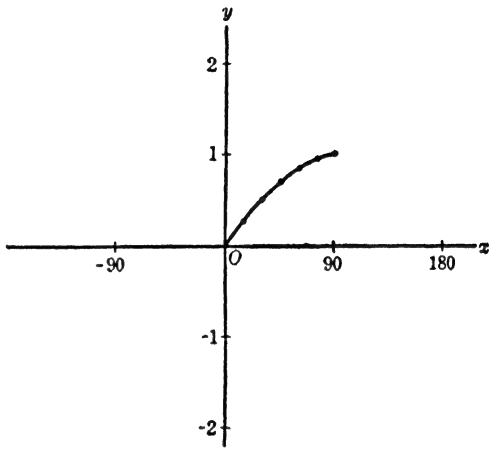
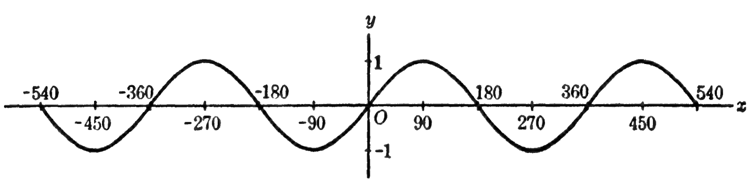


FIG. 202

The fact that $\sin x$ has the same value at intervals of 360° is expressed by saying that the function $y = \sin x$ is **periodic** (with period 360°). Periodic functions are extremely important in a large variety of physical applications. In fact, the earliest scientific observations made by the human race were doubtless certain periodic phenomena like the alternation of day and night, the cycle of the seasons, the recurrent phases of the moon, the recurrent patterns of stars in the sky, etc. Similarly, many modern scientific phenomena are periodic in character, like sound waves, the vibrations of a violin string, the oscillation of a pendulum, etc. The graph of $\sin x$ suggests wave motion, and in fact the trigonometric functions (which are the simplest of all periodic functions) are very useful in the description of such periodic phenomena.

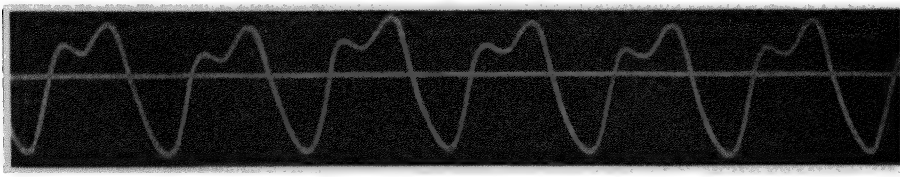


Graph of $y = \sin x$.

FIG. 203

The achievements of indirect measurement are quite remarkable, although they may seem pale beside such wonders as radio,

television, etc. However, the student should remember that he now *understands* the more ancient miracles of trigonometrical indirect measurement, whereas he is doubtless merely *familiar* with the conveniences of the more modern miracles. The elementary applications of the trigonometric functions to the measurement



Photograph of a sound from a violin. An approximate equation of this graph is $y = 151 \sin x - 67 \cos x + 24 \sin 2x + 55 \cos 2x + 27 \sin 3x + 5 \cos 3x$.

FIG. 204

of inaccessible distances discussed in this chapter are interesting and important, but many of the most important and interesting applications of the trigonometric functions of more advanced character are due to the periodicity of these functions. Any book on the advanced physics of sound, light, mechanics, electromagnetics, etc., will be found to make frequent use of the trigonometric functions.

EXERCISES

Plot between 0° and 360° (using 15° intervals) the graph of each of the following. Choose different units on the x and y axes in some convenient manner.

1. $y = \cos x$. **2.** $y = \tan x$. **3.** $y = 2 \cos x$. **4.** $y = \cos 2x$.

5. $y = 2 \cos x + \cos 2x$. (Hint: Use the tabulations of exercises 3 and 4.)

6. $y = \sin 2x + 2 \sin x$.

REFERENCES

F. L. Griffin, *Introduction to Mathematical Analysis*, Houghton Mifflin, Boston, 1921.

L. T. Hogben, *Mathematics for the Million*, Norton, N. Y., 1937.

J. H. Jeans, *Science and Music*, Macmillan, N. Y., 1938.

D. C. Miller, *The Science of Musical Sounds*, Macmillan, N. Y., 1916.

J. Redfield, *Music, a Science and an Art*, Knopf, N. Y., 1928.

C. N. Shuster and F. L. Bedford, *Field Work in Mathematics*, American Book Co., N. Y., 1935.

A. N. Whitehead, *Introduction to Mathematics*, Henry Holt, N. Y., 1911.

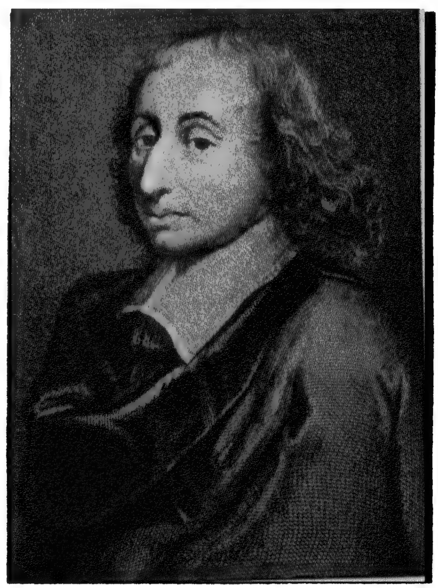
Histories of Mathematics.

Textbooks on Plane and Spherical Trigonometry.

Chapter XIII

PROBABILITY AND STATISTICS

131. Introduction. It is difficult to give a precise definition of statistics. Speaking loosely, statistics is concerned with coming to general conclusions, whose truth is more or less probable, concerning large classes of phenomena, by means of the study of large numbers of observations. It is therefore of the greatest importance in connection with the inductive logic of experimental methods. The use of statistical inference pervades the social and biological as well as the physical sciences. The insurance business, as well as other commercial affairs, makes essential use of statistics. The subject is based on the study of probability. The mathematical theory of probability had its beginning in correspondence between Fermat and Blaise Pascal (1623–1662) concerning certain questions about games of chance which were put to the puritanical Pascal by an aristocratic acquaintance with a penchant for gambling. Pascal, incidentally, was a child prodigy who displayed phenomenal talent in mathematics and, by the age of 16, had done work of such calibre that Descartes is said to have refused to believe it was done by one so young. Pascal made noteworthy contributions to many branches of mathematics.



Blaise Pascal
1623–1662, French

132. Probability. In everyday speech, the word "probable" is used in a very vague sense. For example, we say "it will probably rain tomorrow," "the Yankees will probably win the world series," "*A* probably murdered *B*," "the patient will probably recover." What meaning can be attached to the word "probability"?

Some philosophers have maintained that probability is merely a measure of belief. But such a subjective matter could hardly be of great scientific value. Since different people are notoriously capable of believing opposite statements with equal firmness, whose belief are we to measure? In fact, the same person may very well believe opposite statements at different times. Clearly, a more objective meaning must be given to the word probability if we are to build up a useful mathematical theory about it.

Example 1. Consider the tossing of a coin. There are only two ways it can fall: head or tail. We consider either of these possibilities equally likely. We say that the probability of obtaining a head is 1/2.

This suggests the following definition.

DEFINITION. *If a certain event can occur in a total number t of alternative ways, all of which are regarded as equally likely, and a certain number s of these ways are considered successful or favorable, then the number s/t is called the* **probability** *of success.*

In example 1, $s = 1$, and $t = 2$.

Example 2. Suppose a bag contains 10 balls of which 3 are red, 4 yellow, 2 blue, and 1 green. The probability of drawing a red ball from the bag is 3/10, the probability of drawing a yellow ball is 4/10, the probability of drawing a blue ball is 2/10, and the probability of drawing a green ball is 1/10.

If the event can happen *only* in the way which is regarded as successful, then $s = t$ and the probability of success is 1. In example 2, the probability of drawing a colored ball is 1. If the event can never happen in the successful way, then $s = 0$ and the probability of success is $0/t = 0$. In example 2, the probability of drawing a white ball is zero. Hence a probability of 1 indicates certainty and a probability of 0 indicates impossibility.

The probability of any event will be a number between zero and one, inclusive.

Notice that the assumption that all the possibilities are equally likely is important. We might be tempted to say, in example 2, that the probability of drawing a green ball is 1/2 because there are only two possibilities, namely drawing a green ball, and not drawing a green ball, of which one is favorable. But this would be wrong because these two possibilities are not regarded as equally likely. The correct answer is 1/10.

Example 3. Suppose 3 coins are tossed. What is the probability that we will get exactly two heads?

Now there are eight possibilities, namely: HHH; HHT, HTH, THH; HTT, THT, TTH; TTT. Of these 8, there are three which are successful. Hence the probability is 3/8.

Consider the statement that in tossing a single coin the probability of a head is 1/2. If we knew precisely the size, shape and distribution of weight of the coin and the magnitude and direction of the force applied to the coin in the act of tossing it, and all the other pertinent physical conditions, we could predict, by the theorems of mechanics, how the coin would move and hence whether it would fall head or tail. Therefore, it is sometimes argued that probability is based entirely on ignorance. While it is hard to believe that the extensive practical applications of probability are all based on ignorance, the above argument is not to be laughed off too easily. It has some partial truth in it, although it is essentially due to confusion about the meaning of probability. The confusion arises partly from the mistaken notion that probability tells us something about a particular event. The statement that the probability of a head is 1/2 tells us nothing about any particular toss. It is usually interpreted to mean that if we toss a coin a "sufficiently large" number of times, the event "heads" would occur about half the times; furthermore, we assume that the more often we toss it, the closer the ratio of heads to total number of tosses will be to 1/2. In mathematical language, we assume that if $f(n)$ is the number of favorable events (heads) actually occurring in n trials (tosses) then the "relative frequency" $f(n)/n$ approaches the probability s/t (1/2) as a limit as n increases indefinitely. In symbols we assume that

$\lim\limits_{n\uparrow} \dfrac{f(n)}{n} = \dfrac{s}{t}$; that is, that $\dfrac{f(1)}{1}, \dfrac{f(2)}{2}, \cdots, \dfrac{f(n)}{n}, \cdots \to \dfrac{s}{t}$. Thus probability is interpreted as relative frequency. For example we might find the following results by experimenting with a coin:

n	10	20	50	100	1000	10,000	100,000	$\cdots$
$f(n)$	4	13	23	52	493	5,023	50,117	$\cdots$
$f(n)/n$	.4	.65	.46	.52	.493	.5023	.50117	$\cdots$

However, we might very well find entirely different results experimentally. For instance, we would certainly find a different relative frequency of heads if our coin were weighted so as to fall heads more easily. This is why we said we *assume* that the relative frequency will approach 1/2 as a limit as the number n of trials increases indefinitely. In practice, this assumption can never be put to the test of experience, except approximately, since we do not live long enough to let the number of trials increase *indefinitely*.

A great deal of the confusion concerning the theory of probability is due to simple failure to distinguish between the theory of probability as a branch of pure mathematics (that is, an abstract mathematical science) and the theory of probability as a branch of applied mathematics (that is, a concrete interpretation or application of the abstract mathematical science). We must recall that pure mathematics is not concerned with truth but asserts only that certain theorems follow logically from certain assumptions; while in applied mathematics we are concerned about the truth of our assumptions but we can seldom be certain about it. Now, as a branch of pure mathematics (that is, a deductive science), the theory of probability may be based * on the following assumptions (among others):

1. The alternative possibilities are equally likely.
2. The relative frequency of success will approach the probability (as defined above) as a limit as the number of trials increases indefinitely.

* Many different points of view have been taken toward the foundations of probability.

When we apply the resulting abstract mathematical science to reality in a definite concrete situation, like that of an actual coin, we cannot be absolutely sure that these assumptions are actually satisfied. We have then a concrete interpretation or scientific theory whose actual truth can be verified at best in an approximate sense. As long as the experimental results are concordant with our theoretical predictions we have a good theory. The results of the theory have been borne out by experiment with remarkable success when applied properly. Compare Chapter II, sections 8 and 9.

In the following exercises, the student has only to determine the total number t of possibilities and then the number s of favorable possibilities. This is usually a matter of simple arithmetic. The following principle will be found helpful.

If one event can occur in h different ways, and if, no matter in which of these h ways the first event occurs, a second event can occur in k different ways, then the two events can occur together in $h \cdot k$ different ways. On the other hand, either the first or the second event (but not both) can occur in $h + k$ ways.

Example 4. In one box are 3 tickets marked a, b, c. In a second box are four tickets marked 1, 2, 3, 4. (a) We are to pick one ticket from each box. How many different pairs of symbols, each pair consisting of one letter and one number, can we pick? (b) If we are to pick one ticket from either box, how many possible choices are there?

(a) Clearly, we can pick a letter in 3 ways. No matter which letter we pick, we can pick any one of the 4 numbers. Hence, there are $3 \cdot 4$ or 12 possible pairs: $a1$, $a2$, $a3$, $a4$, $b1$, $b2$, $b3$, $b4$, $c1$, $c2$, $c3$, $c4$. (b) There are $3 + 4$ or 7 possible choices if we are to pick just one ticket.

Example 5. Three coins are tossed. In how many ways can they fall? The first coin can fall in two ways, head or tail. No matter which way this first event occurs, the second coin can then fall in two ways. No matter how these first two events occur, the third coin can then fall in 2 ways. Hence there are $2 \cdot 2 \cdot 2$ or 8 possibilities. They were listed explicitly in example 3, above.

For further work on a priori probability, see Appendix, sections 171–174.

EXERCISES

1. Toss an actual coin 100 times and record the number of heads obtained in 10, 20, 30, $\cdots$, 100 tosses.

2. What is the probability of throwing a "six" with a single die?

3. What is the probability of throwing a total of "two" with a pair of dice?

4. What is the probability of throwing a total of "three" with a pair of dice? A total of "four"? A total of "seven"? Which is more probable, a total of "three" or a total of "seven"?

5. What is the probability of obtaining exactly one head in tossing two coins simultaneously? At least one head? Exactly two heads? No heads?

6. Actually toss two coins 100 times and record the relative frequencies of 0, 1 2 heads.

7. What is the probability of obtaining exactly 1 head in tossing three coins? Two heads? Three heads? No heads? At least one head? At least two heads?

8. What is the probability of obtaining exactly 0, 1, 2, 3, 4 heads in tossing 4 coins? At least one head? At least two heads? At least three heads?

9. What is the probability of picking a diamond from a pack of 52 ordinary playing cards? An ace? The ace of spades?

133. Statistical probability. Probability as defined in section 132 is usually called **a priori probability,** because with this definition we deduce that the probability of obtaining heads in tossing a single coin, for example, is 1/2 from our hypotheses without actually experimenting by tossing the coin. A priori probability is of the greatest importance in both theoretical and practical work. However, one can expect to work out the a priori probability only in examples like those of the preceding section where one has sufficient data to determine the probability without experiment. It would hardly do for such things as life-expectancy in insurance. There is no a priori way in which one can say that a man, 40 years of age, in certain circumstances, has an 89/100 chance of reaching the age of 50. For such matters, we take relative frequency as the definition of probability. That is we would collect data about a large number of men 40 years of age in similar circumstances and actually count the number who reach the age of 50. The total number n is the number of "trials" and the number $f(n)$ who live to the age of 50 is the number of "successes."

The ratio $f(n)/n$ is called the **statistical probability** of success. Statistical probability is understood to be relative to some body of knowledge. With more data, the statistical probability might change. We could build up the theory of probability by taking $\lim_{n\uparrow} \frac{f(n)}{n}$ as the definition of the "true" probability, assuming that this limit exists. Then the successive statistical probabilities $f(n)/n$ as n increases are taken as approximations to the "true" probability. But in practical work we can never ascertain the value of this limit, since we are necessarily forced to use a finite value of n. If we make more trials so that the total number of trials is n' ($> n$), then $f(n')/n'$ may be considered a better approximation of the "true" probability than $f(n)/n$. This is, in fact, what is done in connection with life-expectancy tables in insurance and other statistical data. When more trials are made or more data gathered the statistical probability, of life-expectancy for example, is changed if necessary. Statistical probability is called also **a posteriori, empirical** or **experimental probability.**

It would be foolish to suppose that statistical probability is independent of assumptions because it is based on actual observation. We are still assuming among other things that the sequence $f(1)/1, f(2)/2, f(3)/3, \cdots, f(n)/n, \cdots$ approaches a limit as n increases indefinitely and that for a given "sufficiently large" value of n (whatever that may be) the fraction $f(n)/n$ is really a good approximation of this limit.

While we can never be sure that our assumptions are really true in a concrete situation, nevertheless the theory of probability has been found to work very well in at least an approximate sense. The situation is quite the same as in the case of Euclidean geometry or Newtonian mechanics, or any other (applied) scientific theory (see Chapter II, sections 8 and 9). We are simply faced again with the recurring distinction between pure mathematics and its application to reality. The ultimate test of the applicability of a theory to reality is whether or not it works. The theory of probability has been found to be of tremendous value in such diverse subjects as physical chemistry, statistical mechanics, heredity, biometrics, econometrics, the social sciences in general, the theory of gases, insurance, games of chance, experimental method in general, etc.

Example. Out of 100,000 ten year old children, about 70,000 reach the age of 50 and about 58,000 reach the age of 60. On the basis of these data, we would say that the probability that (*a*) a ten year old will live to the age of 50 is 70/100; (*b*) the probability that a ten year old will reach the age of 60 is 58/100; and (*c*) the probability that a 50 year old will reach the age of 60 is 58,000/70,000 or nearly 83/100. Needless to say, the assertion that a 50 year old has, on the basis of these data, an 83% chance of reaching the age of 60 means absolutely nothing to any particular 50 year old since it takes no account of his health or other circumstances. But the statement has a fairly reliable meaning for large groups, as may be seen from the soundness of the insurance business which is based upon just such principles of statistical probability.

EXERCISES

1. Of 100,000 ten year olds, it is found that 74,000 reach the age of 45 and 65,000 reach the age of 55. Find the probability, on the basis of these data, that (*a*) a ten year old will reach the age of 45; (*b*) a ten year old will reach 55; (*c*) a 45 year old will reach 55.

134. Statistics. While we could go a little further into the study of probability and statistics here,* we could not go very far without a more intensive study of certain technical parts of mathematics. In fact, many mistaken conclusions are drawn from statistical data by people whose mathematical training is insufficient. It is not because they make mistakes in calculation but because they often do not understand the mathematical (logical) basis for the formulas they use. If all the people who publish statements in newspapers and magazines, beginning with the words "Statistics prove . . ." were laid end to end, there would probably be far less pseudo-scientific superstition among educated laymen. Abuses of this kind have led to the well known classification of lies into 3 kinds: lies, worse lies, and statistics. This classification does not refer to statistics properly used. Although we cannot develop the theory of statistics logically in such a non-technical book as this, we can discuss some of its simpler ideas and point out some of the more glaring, cruder,

* See Appendix, sections 171–174.

elementary follies that are often committed. This we shall do in the remainder of the chapter.

135. Accuracy. Many people who ought to know better make the mistake of thinking that a statistical report has a high degree of accuracy if its results are carried out to a great many decimal places. To see how ridiculous this notion is, let us consider the following experiment. Take a large tub of water and place seven empty pails on the floor, being careful to roll up the rug first. Now spill the water from the tub into the seven pails, giving each pail an equal share as nearly as you can. Suppose now you say that each pail contains roughly 1/7 of a tub of water. This does not look very accurate. If, however, you divide 1 by 7 and get .142857 ⋯ of a tub your report looks like a masterpiece of scientific accuracy, carried out to 6 decimal places. However, you could just as well have carried out the division of 1 by 7 to 18 decimal places, obtaining .142857142857142857 ⋯ of a tub without improving the actual accuracy at all. You really know no more than you did when you said roughly 1/7 of a tub. *Genuine accuracy depends on the accuracy of your data.* Thus in a table of logarithms or trigonometric functions, more decimal places really give more accuracy, but this is because of the source of the additional decimal places. The value of π is another example. The true value is $\pi = 3.14159265 \cdots$. Taking more decimal places gives greater accuracy. But if you take $\pi = 22/7$, roughly, you may get a decimal expression $3.14285 \cdots$ by division. Now if you take more than the first two decimal places from the latter expression, your increase in accuracy will be illusory because only the first two decimal places are correct. In fact, you will really be departing further and further from the true value given above. The italicized statement above was not appreciated by the visitor to the museum who said, "This fossil is 5,000,003 years old because when I was here 3 years ago the guide told me it was 5,000,000 years old." Nor is it understood by the trigonometry student who insists that the height of a mountain is 5,253.67952 ft.

136. Index numbers. In making comparisons with respect to qualitative considerations, it is often convenient to sharpen vague comparisons by using numerical indices of some sort. Now

numbers are used sometimes as direct measurements of quantitative characteristics and sometimes merely as identification labels or names; but in the case of qualitative comparisons they are usually used to indicate the relative order of the objects studied with respect to the given quality. Thus the intelligence quotient (I.Q.) of a person is essentially his score on a certain test. Similarly economists use various numerical scores or indices. Now in the interpretation of these scores, it must be remembered that we can usually use them only in so far as they place our specimens in an order. Thus if three people get I.Q.'s of 150, 100, and 75 respectively, we say that (as far as this test is concerned) the first is more intelligent than the second who is more intelligent than the third. But it is silly to say that the first is twice as intelligent as the third on the ground that $150 = 2 \cdot 75$. Similarly, to take an extremely simple case, if we arrange a class of boys according to size beginning with the smallest and assign to each one an index number indicating his place in line, we could not conclude that the boy with index number 20 is twice as tall as the boy with index number 10 merely because $20 = 2 \cdot 10$.

137. Correlation. Suppose we were to find that the marks of 100 students in intermediate algebra and trigonometry are as in the adjoining table. We might get the impression that there is some connection or relation between the algebra marks x_1, x_2, x_3, $\cdots$ of these students and their trigonometry marks y_1, y_2, y_3, $\cdots$. Thus, we observe that, by and large, most of those who did well in algebra did well in trigonometry and most of those who did poorly in algebra did poorly in trigonometry. Of course the conclusion that a student's mark in trigonometry will be much like his mark in algebra is not justified if we apply it to some particular student. But it does seem as

Student	x = *mark in int. alg.*	y = *mark in trig.*
1. Ames	$x_1 = 60$	$y_1 = 50$
2. Brown	$x_2 = 85$	$y_2 = 85$
3. Camp	$x_3 = 90$	$y_3 = 95$
4. Davis	$x_4 = 60$	$y_4 = 60$
5. Ellis	$x_5 = 60$	$y_5 = 55$
6. Fischer	$x_6 = 75$	$y_6 = 65$
7. Green	$x_7 = 75$	$y_7 = 85$
8. Harris	$x_8 = 70$	$y_8 = 60$
9. Irving	$x_9 = 70$	$y_9 = 75$
10. Jones	$x_{10} = 80$	$y_{10} = 90$
11. Kirk	$x_{11} = 80$	$y_{11} = 50$
12. Lewis	$x_{12} = 60$	$y_{12} = 90$
⋮	⋮	⋮

though there is some approximate relationship between the values of the variable x and the values of the variable y. It is clear from the table that y is not a single-valued function of x, for the same value of x has associated with it many values of y. For example, the first, fourth, fifth and twelfth students all have $x = 60$ but they have different values of y. However, it does seem as though, in general, most of the values of y may be fairly well approximated by a single-valued function of x. Let us see what this means by plotting on a graph the points whose coordinates are (x_1,y_1), (x_2,y_2), (x_3,y_3), $\cdots$. This gives us what is often called a **scatter diagram** (Fig. 205). These points are roughly clustered about a straight line (drawn on Fig. 205) although there are some stray points which are quite far from the line. Thus we would say that the values of y are fairly well approximated by a linear function of x; in technical terms, we say that the variable y has a high degree of **simple correlation** with the variable x. The case where the values of y are fairly well approximated by a linear function of x is also called **linear correlation** and it is possible to set up a certain expression called the **coefficient of correlation** which measures how closely the scatter diagram approximates a linear function. This coefficient is so constructed that it is positive if the y's generally increase as the x's increase (as in Fig. 205) and negative if the y's generally decrease as the x's increase (as in Fig. 206). If all the y's really lay along a straight

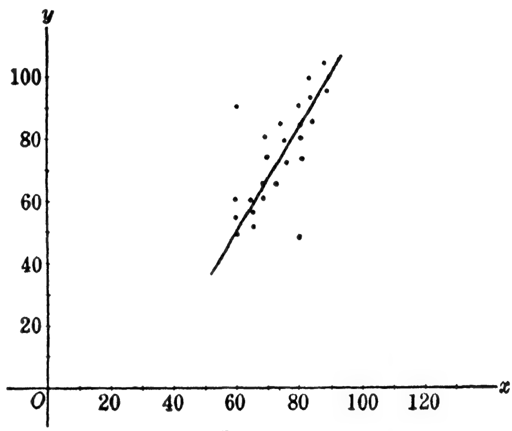


FIG. 205

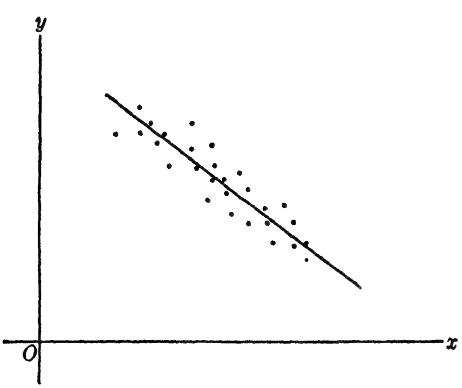


FIG. 206

line the correlation would be perfect. There are other types of correlation than the usual simple or linear type. The idea is that if the values of the variable y are, in general, fairly well approximated by a single-valued function of x, then y is said to be highly correlated with x. Thus Fig. 207 illustrates a scatter-diagram exhibiting a high degree of correlation which is not of the linear type. We shall make no attempt to give a precise definition of correlation or of any of the coefficients of correlation in general use. We shall leave the idea of correlation on an intuitive level and hope that the sketchy, non-technical explanation given here conveys the idea to the reader, at least roughly.

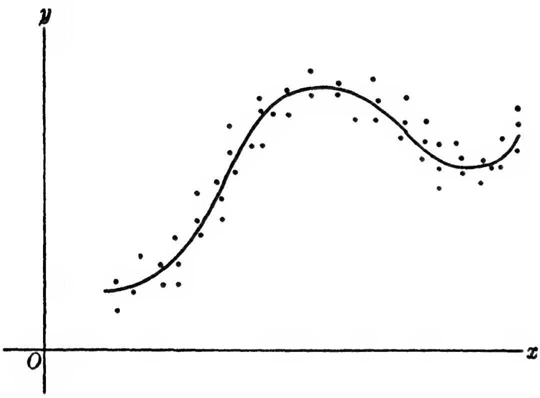


FIG. 207

When two variables have a high degree of correlation, we might be inclined to suspect some genuine relationship between them. Thus in the case above, we might suspect that poor work in intermediate algebra often causes poor work in trigonometry. Such a conclusion must not be rashly drawn. A high degree of correlation may, indeed, be indicative of such a causal relationship, if supported by other evidence, but can never justify such a conclusion by itself. We might find that the annual numbers of births in Chicago have a high correlation with the corresponding annual numbers of storks observed in Central Asia, but we should be cautious about asserting that the high correlation implies a causal relationship. In the absence of additional evidence, such a correlation might well be accidental. To cite a somewhat different example, we might observe a high correlation between the daily numbers of people at nearby beach resorts and the corresponding daily numbers of people taking boat rides on the Hudson River. It would not be correct to say that, because these variables have a high correlation, one causes the other. Other evidence might indicate that both of these variables are dependent upon the weather. However, a high correlation may indi-

cate the *possibility* of some causal relationship between the two variables, and it is often used for that purpose. But the causal relationship cannot be inferred from the correlation alone. For example, if we measure the lengths of an iron bar at various temperatures we might find that the length increases and decreases with the temperature. This might lead us to suspect that a change in temperature will cause a change in the length of the bar. Such methods are very much used in economics, biology, education, medicine, etc., and are very valuable tools in inductive reasoning by means of what is called the method of concomitant variation. But caution is necessary in drawing conclusions from statistical correlation. It is also possible to study correlation involving more than two variables.

138. Sampling. Often statistical information about large numbers of people, say the population of the United States, is obtained by gathering data from a "random sample" or "typical cross-section" of the population. Naturally, the greatest care must be taken to insure that the sample shall really be "random" or "typical" and not biased in some direction pertinent to the investigation in hand. Thus a poll on the desirability of Sunday Blue Laws might conceivably yield biased and different results if the sample of people questioned were chosen largely from the crowds emerging from the churches on Sunday or from the crowds emerging from a ball game on Sunday. A really "typical" sample should include the same proportion of people holding various views as is found in the entire population; that is, it should be truly representative. It is not easy to decide how to obtain a random sample. The technique used by a certain commercial organization engaged in such polls is roughly to choose a small number of people in such a way that the proportion of Republicans, Democrats and other political parties, the proportion of urban and rural residents, the proportion of old and young, the proportion of white and colored, the proportion of workers in various fields, the proportion of those living in various geographical divisions, and so on, are the same as in the entire population as revealed by the census. This method has enabled them to make very good predictions. The possibility must be borne in mind, however, that a sample which has shown itself to be typical or representative on one question, may by chance be unrepresenta-

tive on another. Even with the most careful methods of sampling, conclusions cannot be regarded as certain. In any case, any such conclusions must be interpreted in the light of how the sample was chosen. It is worth noting that large numbers of individuals are not a necessity in obtaining a good sample. One individual may be sufficient if it is a truly representative individual. For example, a chemist is satisfied that a reaction will always take place under given conditions after having performed the experiment once, if he is fairly sure his samples have been typical or representative of the substances involved.

139. Averages. In dealing with large collections of numbers (measurements or statistical data of any kind) it is often humanly impossible to draw quantitative conclusions about the group as a whole from the *entire* collection of individual numbers as they stand, simply because there are too many of them. Therefore, we would like somehow to describe the general quantitative characteristics of the entire group, at least in certain respects, by means of a few numbers. Two kinds of numbers are often used for this purpose. One kind, called *averages*, are for the purpose of describing the "central tendency" of the collection. Thus the collection of numbers 40, 45, 48, 50, 50, 50, 50, 50, 52, 55, 60 has the "average" 50; that is, 50 is the value about which the entire collection "centers." Another kind of numbers, called *measures of dispersion*, are for the purpose of describing how the numbers of the collection deviate from some average or central value. The collection of numbers 10, 30, 40, 50, 50, 50, 50, 50, 60, 70, 90, has the average 50 just as in the example above; but the numbers of this collection are dispersed very differently about this central value. We shall not discuss measures of dispersion in great detail here but shall examine more closely some of the measures which are often used as "averages." There are many such measures and we shall use different names for them to avoid the indefiniteness that usually accompanies the word "average."

The **arithmetic mean** of a set of n numbers $x_1, x_2, x_3, \cdots, x_n$ is their sum divided by their number; that is,

$$\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}.$$

Example 1. If a student gets marks of 70, 80, 90, 100 during the term, his "average" (arithmetic mean) is 85. Note that the arithmetic mean need not be one of his marks.

Example 2. A factory owner pays the following salaries:

Position	*No. of men*	*Weekly salary*
manager	1	$200
foremen	2	40
skilled workers	10	30
unskilled workers	11	20
apprentices	1	10

Clearly, the arithmetic mean of the numbers in the salary scale at the right, namely $\dfrac{200 + 40 + 30 + 20 + 10}{5} = 60$ would give a false impression of working conditions. Even the owner would doubtless be unwilling to say that the "average" salary paid in his factory is $60 per week. He might, however, say that the "average" salary is

$$\frac{1 \cdot 200 + 2 \cdot 40 + 10 \cdot 30 + 11 \cdot 20 + 1 \cdot 10}{25} = 32.40$$

or $32.40 per week. This is called the **weighted arithmetic mean** since each number in the salary scale is weighted according to its importance, the importance being measured by the number of people receiving that salary.

Example 3. A student receives a mark of 90 in a 5-credit course, 85 in a 4-credit course, and 65 in a 1-credit course. If his "average" were computed by the arithmetic mean, it would be $(90 + 85 + 65)/3 = 80$. But the student might be considered to have just cause for complaint. If the number of credits may be taken as indicating the relative importance of the courses, it would be fairer to use the weighted arithmetic mean which would give an average of

$$\frac{5 \cdot 90 + 4 \cdot 85 + 1 \cdot 65}{10} = 85.5.$$

Arithmetic means have the property that they give great importance to extremely large or extremely small numbers. Thus, in example 2, if the owner announced that the average salary was the weighted arithmetic mean \$32.40 per week, the labor union might object because the extremely large salary paid to the manager (who happened to be the owner's son-in-law) exerted an undue influence on the result. If the manager's salary is left out, the average (weighted arithmetic mean) salary is \$25.42 per week.

In taking measurements in a laboratory experiment, it is natural to suppose that extremely wrong measurements do not occur. Thus if a scientist takes a measurement a large number of times, obtaining only slightly different results each time, it is usual to regard the arithmetic mean of the various readings as the "most probable" value of the measurement. A great deal of useful statistical theory is based on this assumption.

The **mode** is the number, in a collection of numbers, which occurs most frequently, if such a number exists. In example 2, the modal salary is \$20. The mode has the property that it is not influenced by isolated extreme values, like the manager's salary. But it has certain defects. For instance, the frequency with which various numbers occur may be the same or very much the same, so that the mode becomes meaningless or useless. For example, if a student gets marks of 90, 90, 80, 80, 75, 75, 65, 65, the frequency of each mark is 2 and there is no modal mark. Suppose a second student has marks of 100, 99, 98, 90, 90, 85, 85, 85, 65, 65, 65, 65. The modal value is 65 but it is hardly representative of the student's work. In fact, if the marks were classified by intervals of 10 the student's marks would have the following frequencies

interval	100–90	89–80	79–70	69–60
frequency	5	3	0	4

and the modal interval would be 100–90. If we used intervals of 5, however, we would have

interval	100–95	94–90	89–85	84–80	79–75	74–70	69–65
frequency	3	2	3	0	0	0	4

and the modal interval would be 69–65. Statements about "the average man" usually refer to the modal man.

If a collection of numbers is arranged in order of magnitude the term in the middle position is called the **median,** if such a number exists. If the number of numbers in the collection is odd, the median always exists. If the number of numbers in the collection is even, there are *two* middle terms and we usually call the arithmetic mean of these two terms the median. Thus in the case of the first student just referred to, the median mark is 77.5. The second student has a median mark of 85. In example 2, the median salary is \$30. *All the various types of "averages" yield, in example 2, entirely different results.* It is possible that if the employer were announcing the "average" salary he would use the largest value while the labor union would announce the smallest value. Therefore statements about "averages" must be carefully interpreted in the light of how the "average" was calculated and whether the type of average used was appropriate to the given situation.

Remark. There are other useful expressions which may be called averages. Suppose an investment dwindles in value to 20% of the original value during the first year, and then dwindles during the second year to 80% of the value it had at the beginning of the second year. What steady rate of decrease would yield the same resulting value at the end of two years? If r is the steady rate, we would have $r \cdot r = \frac{20}{100} \cdot \frac{80}{100}$ or $r^2 = \frac{1600}{10000}$. Hence $r = \sqrt{\frac{1600}{10000}} = \frac{40}{100} = 40\%$. This is called the geometric mean of the two given rates. In general, the **geometric mean** of two given quantities a and b is $\sqrt{ab}$. The **geometric mean** of n given quantities $x_1, x_2, \cdots, x_n$ is $\sqrt[n]{x_1 x_2 \cdots x_n}$.

The collection of numbers 10, 30, 40, 50, 50, 50, 50, 60, 70, 90 has the arithmetic mean 50. The collection of numbers 40, 45, 48, 49, 50, 50, 51, 52, 55, 60 also has the arithmetic mean 50. In fact the median and mode for each of these collections is also 50. Hence both exhibit the same "central tendency." That is, both collections center about the number 50. But clearly the numbers of the first collection are dispersed about the number 50

in quite a different manner from that in which those of the second collection are dispersed about the number 50. To describe such differences we use various "measures of dispersion" of which we shall mention only one here. Each number in the collection differs from the arithmetic mean by some amount, called its **deviation from the mean.** The deviations from the mean of the numbers in the first collection above are 40, 20, 10, 0, 0, 0, 0, 10, 20, 40. The arithmetic mean of these deviations is called the **average (or mean) deviation from the mean.** In this case it is 14. In the second collection above, the deviations from the mean are 10, 5, 2, 1, 0, 0, 1, 2, 5, 10. The average (mean) deviation from the mean in this case is 3.6. This indicates that the numbers in the second collection are not as widely dispersed about the mean value as those of the first collection.

EXERCISES

1. A student's marks in 10 subjects are 95, 90, 85, 85, 85, 78, 77, 73, 72, 60.

(*a*) What is the arithmetic mean of his marks?

(*b*) What is the median mark?

(*c*) The modal mark?

(*d*) Using the intervals 95–100, 90–94, 85–89, · · ·, 60–64, what is the modal interval?

(*e*) Using the intervals 90–100, 80–89, 70–79, 60–69, what is the modal interval?

(*f*) If the first 5 marks are in major subjects counting 3 credits each and the last 5 are in minor subjects counting 1 credit each, find the weighted arithmetic mean, taking the number of credits as the measure of relative importance.

(*g*) Find the average (mean) deviation from the mean.

2. The salaries in a certain firm are as follows:

Position	*No. of men*	*Weekly salary*
foremen	3	50
skilled workers	20	40
unskilled workers	19	30
apprentices	8	10

Find (*a*) the modal salary; (*b*) the median salary; (*c*) the arithmetic mean of the salaries in the right hand column; (*d*) the weighted arithmetic mean, the importance or weight of each salary being taken as the number of men receiving it.

3. A student's marks in 10 subjects are 98, 90, 93, 97, 85, 87, 88, 67, 70, 65. Using these data, answer the same questions as in exercise 1.

140. Normal frequency distributions. If we toss 3 coins there are 4 possible results: no heads, one head, two heads, or three heads. The a priori probability of these results would be 1/8, 3/8, 3/8, 1/8, respectively. These are the expected relative frequencies with which these results would occur if we made a large number of actual trials. We could represent this theoretical rela-

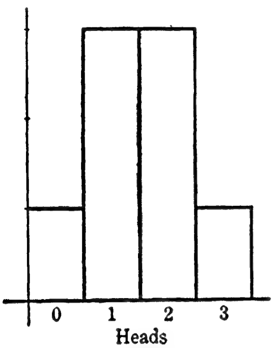


FIG. 208

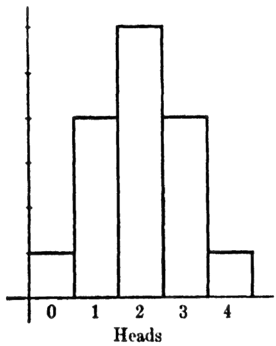


FIG. 209

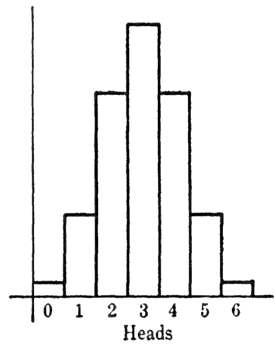


FIG. 210

tive frequency distribution by a "step graph" as in Fig. 208. If we toss 4 coins, the a priori probabilities of 0, 1, 2, 3, 4 heads are respectively 1/16, 4/16, 6/16, 4/16, 1/16. These theoretical frequencies would yield a step graph like that in Fig. 209. If we toss 6 coins, we get a step graph like that in Fig. 210. As the number of coins increases and the width of each rectangle in the step graph and the unit on the vertical axis are decreased accordingly (Fig. 211), the step graph approaches a smooth curve having the general shape of a bell (Fig. 212). It can be proved that this curve has an equation of the form $y = ce^{-hx^2}$ where c and h are constants and e is the number $2.71828 \cdots$ discussed in section 110. This is the so-called normal probability curve or normal frequency distribution. If we make a measurement in the laboratory a large number n of times and we assume (among other things) that the arithmetic mean of our readings is the most probable value, it can be shown that the frequency distribution of our errors will approach a curve of this shape as n increases, and that the area under the curve for a small interval Δx will be proportional to the probability p that an error will be in that interval; that is, the area of this strip divided by the

total area under the curve will be p. We cannot go into the details of this curve here except to say that the theorem that a frequency distribution will have approximately that shape is based on many assumptions. It is assumed that the number of cases is very large. It is assumed also that all deviations are due to chance and not to some special one-sided cause. It is silly to attempt to *force* any set of frequencies to fit this bell curve if they do not do so by themselves. For example, to arrange the marks of a class of thirty students so as to conform with the relative frequencies of a bell curve is sheer absurdity. Even if a "very large" number of measurements do not fit a bell curve we have no justification for trying to force them to fit the curve. For example, if a school desires to give 10% of its students a failing grade, 10% a grade of A, and so on, conforming with a bell curve, it certainly may do so; but it should be realized that this amounts to an arbitrary placement of levels of attainment for the various grades in such a position as will yield the desired frequencies. It is not correct to say that this procedure is "logically justified" by the mathematical theory of statistics unless one is sure that the concrete situation we are dealing with actually satisfies all the assumptions on which this part of the theory of statistics is based. A frequency distribution may well look quite different from the so-called "normal" bell curve.

EXERCISES

1. Toss 4 coins 100 times and compare the actual relative frequencies of 0, 1, 2, 3, 4 heads with the theoretical frequencies given in section 140. Make a step graph like that of Fig. 209 for the actually observed relative frequencies.

2. What would be the theoretical relative frequencies (a priori probabilities) for 0, 1, 2, 3, 4, 5 heads in tossing 5 coins simultaneously? Make a step-graph for

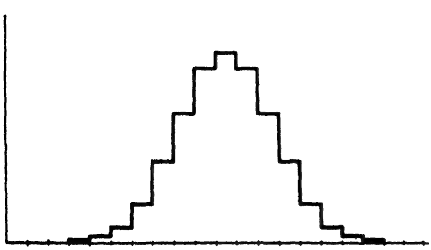


FIG. 211

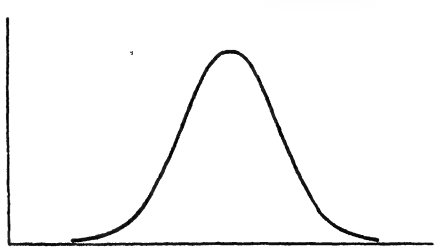


FIG. 212

these probabilities. Actually toss 5 coins 100 times. Make a step-graph for the observed relative frequencies.

3. A certain regiment has as a requirement for membership the stipulation that one must be at least 6 ft. tall. If a frequency curve is made for the heights of the members of this regiment, the curve looks not like a bell-curve but like that of Fig. 213. Why?

4. A college requires for entrance an average mark of 80% for 4 years of high school. Would a frequency curve of the high school averages of its entering class be more likely to look like a bell-curve or like the curve of Fig. 213? Why?

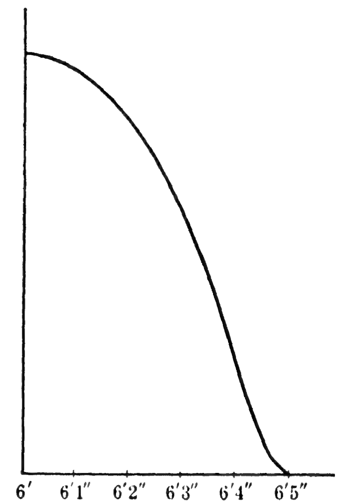


FIG. 213

141. Curve-fitting. Suppose a definite number of observations are tabulated which relate two variables x and y. It is often desirable to approximate the functional relationship given by the table by some simple function like $y = f(x)$ where $f(x)$ is a polynomial, say. That is, plotting the points (x,y) given by the table, we want a curve $y = f(x)$ that

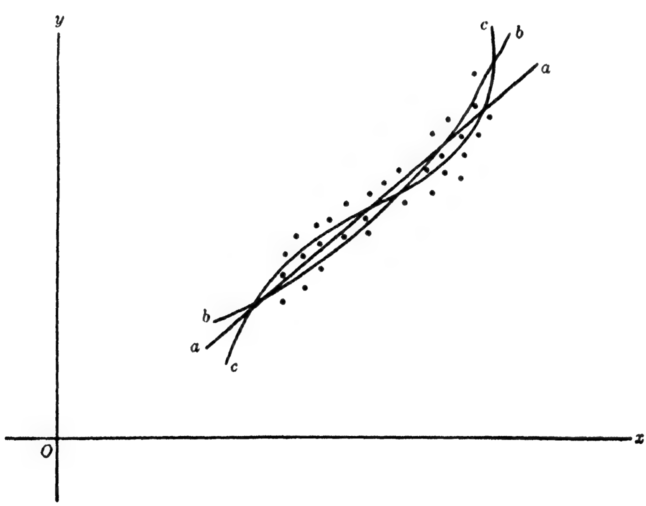


FIG. 214

approximately fits those points. Now, it is intuitively obvious that many different polynomials may approximate the plotted points (Fig. 214). Thus we might choose a straight line, (*aa*, Fig. 214), for the sake of simplicity, as the graph of the function to be

used as an approximation to the plotted points. But a curve of degree two (*bb*, Fig. 214), three (*cc*, Fig. 214), or more might fit even better. In fact, even if we decide that a curve of some definite degree is to be used, there may be many of these which fit the plotted points very well. Thus fitting a curve to a finite number of observations should be accompanied by suitable caution and reservations, especially if the resulting curve or function is to be used for interpolation or extrapolation (prediction). (Compare sections 94 and 137.) Various devices are studied in statistical theory, such as the "method of least squares," which are designed to minimize the error committed in fitting a curve of definite type to such data; that is, to choose that one of all curves of a definitely chosen degree which fits the data "best." But the logical basis for such methods, as in the rest of statistics, has to be understood before the methods can be used intelligently. The ultimate test of the value of a fitted curve for the purposes of prediction is the test of experience. Just as in the case of a scientific theory, as long as the curve works well, we may use it; but, if it ceases to work, it must be abandoned or modified.

142. Conclusion. It must not be inferred from this chapter that the theory of statistics is worthless. It is, when properly used, an extremely important and useful branch of mathematics. It is of great value as an aid to the inductive reasoning used in experimental sciences, whether physical, chemical, biological, or social. Its value in insurance, and such commercial affairs, is obvious. What we have meant to point out is that statistical methods are easily misused, especially in the hands of people whose understanding of the mathematical (logical) foundations of the subject is not sufficient to enable them to interpret the results of their calculations properly. For this reason, our belief in statements beginning with "Statistics prove . . .." must be conditioned by our knowledge of the reliability of the person who makes the statement, or still better, by first hand examination of the facts.

EXERCISES

1. The number of people on relief in a certain town has increased steadily in the past few years. Does this mean that business has been steadily getting worse in that town?

2. The ratio of people on relief to the total population in a certain town has increased steadily in the past few years. Does *this* mean that business has been steadily getting worse in that town? Can you suggest another explanation?

3. The ratio of people convicted of crimes to the total population in a certain town has shown a marked rise in the past 50 years. Does this mean that there is more crime than formerly? Can you suggest another explanation?

4. A newspaper editorial says that business has been improving for the past few years because the annual increases in the number of unemployed have been steadily growing smaller. Criticize.

5. The number of people receiving treatments for neuroses has increased rapidly in the past ten years. Does this mean that there are more neurotics now than formerly? Discuss.

REFERENCES

R. G. D. Allen, *Mathematical Analysis for Economists*, Macmillan, N. Y., 1939.

B. H. Camp, *Mathematical Part of Statistics*, Heath, Boston, 1934.

M. R. Cohen & E. Nagel, *Introduction to Logic and Scientific Method*, Harcourt, Brace, N. Y., 1934.

J. L. Coolidge, *Introduction to Mathematical Probability*, Clarendon Press, Oxford, 1925.

H. T. Davis and W. Nelson, *Elements of Statistics*, Principia Press, Bloomington, Indiana, 1935.

W. P. Elderton, *Frequency Curves and Correlation*, Cambridge Univ. Press, Cambridge, 1938.

W. P. Elderton and E. M. Elderton, *Primer of Statistics*, Macmillan, London, 1927.

W. M. Feldman, *Biomathematics*, Griffin & Co., London, 1923.

A. Fischer, *Mathematical Theory of Probabilities*, Macmillan, N. Y., 1922.

T. C. Fry, *Probability and Its Engineering Uses*, van Nostrand, N. Y., 1929.

J. F. Kenney, *Mathematics of Statistics*, van Nostrand, N. Y., 1939.

J. M. Keynes, *Treatise on Probability*, Macmillan, London, 1921.

R. von Mises, *Probability, Statistics, and Truth*, Macmillan, N. Y., 1939.

R. Pearl, *Introduction to Medical Biometry and Statistics*, Saunders, 1930.

C. S. Peirce, *Chance, Love, and Logic*, Harcourt Brace, N. Y., 1923.

H. L. Rietz, A. R. Crathorne, & J. C. Rietz, *Mathematics of Finance*, Henry Holt, N. Y., 1921.

H. Secrist, *An Introduction to Statistical Methods*, Macmillan, N. Y., 1925.

W. A. Whitworth, *Choice and Chance*, G. E. Stechert, N. Y., 1934.

K. P. Williams, *Mathematical Theory of Finance*, Macmillan, N. Y., 1935.

Chapter XIV

NATURAL NUMBERS AND MATHEMATICAL INDUCTION

143. Postulates for the natural numbers. In Chapter V, we showed that algebra could be considered as an abstract mathematical science based on certain postulates, some of which were stated in section 28, in which "number" was an undefined term. In Chapters III and IV, we indicated how the various kinds of numbers could be defined and studied on the basis of the natural numbers alone. Thus in sections 12 and 13, we assumed that we knew what was meant by "natural number" and assumed such things as the addition tables and the associative, commutative and distributive laws, which we regarded as being derived from experience, and in section 16 we defined fractions in terms of natural numbers. Now in any abstract logical system we are necessarily forced to base everything on unproved postulates and undefined terms. Thus, we might convert our work in Chapters III and IV into an abstract mathematical science by taking "natural number" as an undefined term and taking as postulates essentially what we assumed there. But such a treatment would not seem elegant to a mathematician. For both aesthetic and logical reasons, we would like to assume as little as possible. In Chapter III we assumed many things that we could have proved. For example, we could have proved as theorems the addition and multiplication tables if we had begun with the postulates given in the present section, which are due to G. Peano (1858–1932, Italian). This would push the foundation of our algebra still further back to a more primitive level (more primitive in a logical sense, more sophisticated in a psychological sense).

Let us recall that the choice of undefined terms and unproved postulates, on which we base our abstract mathematical science, is largely suggested by experience. Now our experience with counting suggests that every natural number has a successor,

that is, the next natural number in order of magnitude. For example, the successor of 3 would be 4. (Note that we could not say this for all fractions; for example, there is no next fraction in order of magnitude after 1/2.) Thus Peano begins his abstract mathematical science by taking as undefined terms the terms "natural number," "successor," and "1." He then bases the whole system on five postulates or assumptions. The first two are:

P_1. *1 is a natural number.*

P_2. *If x is any natural number, there is, corresponding to it, another natural number called the successor of x, denoted by x'.*

Using these two postulates alone we can proceed as follows. By P_1, 1 is a natural number. By P_2, 1 must have a successor $1'$. Call $1'$ by the name 2. Now 2 is a natural number, and by P_2, it must have a successor $2'$. Call $2'$ by the name 3. And so on.

Notice that these postulates have other concrete interpretations besides the one we are thinking of. For example, suppose we let the undefined terms "natural number" and "successor of" mean "person" and "mother of," respectively, and let "1" mean "yourself." Then P_1 says "you are a person." P_2 says "if x is any person, there is, corresponding to that person, another person called the mother of x." *

The next of Peano's postulates is:

P_3. *1 is not the successor of any natural number.*

This means intuitively that 1 is the *first* of the natural numbers. The next axiom is:

P_4. *If $x' = y'$ then $x = y$.*

Intuitively this means that if two natural numbers are known to have the same successor, they must be equal to each other. Or, to put it another way, two different natural numbers cannot have the same successor. Or, two different natural numbers must have different successors. Note that this postulate is not satisfied by the mother-person interpretation above; for two different persons may have the same mother. The last axiom, called the **axiom of mathematical induction** requires very careful reading:

P_5. *If S is any set of natural numbers which is known to have the following two properties,*

* A somewhat stilted version of the familiar Mother's Day slogan: "Everybody has a mother."

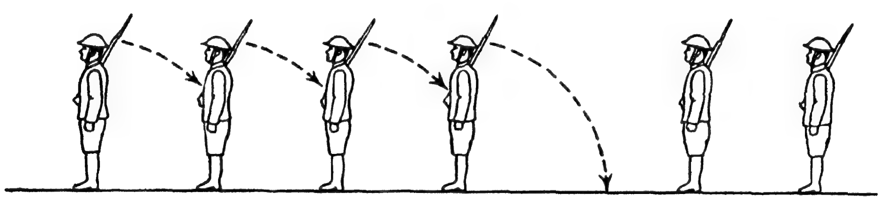
(*1*) *The natural number 1 is in the set S,*

(*2*) *If any natural number k is in the set S, then the successor k' of k must also be in the set S,*

then *all natural numbers are in the set S.*

This postulate seems very plausible. For if 1 is in S then hypothesis (2) says that $1'$ or 2 is in S. Then, since 2 is in S, (2) says that $2'$ or 3 is in S. And so on. The reason that we have to take P_5 as an assumption is precisely because of the words "and so on." We can't very well go on so, for very long, because life is too short and the succession of natural numbers goes on forever. Thus if we kept on with the argument we just started we might, in 3 score and 10 years, get to the point of saying that since 2,345,678,901 is in S, (2) says that $2{,}345{,}678{,}901'$ or 2,345,678,902 is in S. But still we could not assert that *all* natural numbers are in S. However, this assertion does seem plausible; hence we take it as our assumption P_5.

The postulate P_5 may be understood intuitively by the following analogy. Consider an endless single file of tin soldiers beginning somewhere and continuing forever. Suppose we want to make sure that *all* these soldiers fall. It wouldn't do to knock down the first soldier, and then knock down the second, and so on, because we could in this way dispose of only a finite (limited) number of soldiers and this is an *endless* file of soldiers. However, we would be sure they would all fall if we could be sure of two things:



Condition (b) not satisfied

FIG. 215

(a) the first soldier is knocked down;

(b) the soldiers are so spaced that, if any soldier falls, he automatically knocks over the next soldier (his successor).

Condition (a) corresponds to (1) in P_5 and condition (b) to (2)

in P_5. In the case of the tin soldiers, let S be the set of all tin soldiers who must fall. Hence P_5 may be paraphrased to read: *if* S is a set of tin soldiers which is known to have the properties:

(1) the first soldier is in the set S,

(2) if any soldier is in the set S then his successor (the next soldier) must also be in the set S,

then all the soldiers are in the set S.

If we were to base our study of natural numbers on Peano's postulates alone, we would have to define what we mean by addition and multiplication, since nothing is said about these operations in the postulates and they were not taken as undefined terms. Thus we might define $k + 1 = k'$, the successor of k; $k + 2 = (k + 1)'$, the successor of $k + 1$; $k + 3 = (k + 2)'$, the successor of $k + 2$; and so on, where k is any natural number. We would have to prove then, as theorems, that $2 + 2 = 4$, and the rest of the addition table. For example, we could prove $2 + 2 = 4$ as follows. By the definition of addition suggested above, $2 + 2 = (2 + 1)'$, the successor of $2 + 1$. But $2 + 1 = 2' = 3$. Hence, by substitution, $2 + 2 = 3' = 4$. Similarly $3 + 2 = (3 + 1)' = (3')' = 4' = 5$. And so on. We would have to prove, also, the associative, commutative, and distributive laws, etc., all of which we assumed in Chapter III. This can be done, but we shall not attempt it here.

Peano's postulates are essentially an abstract formulation of the process of counting.

We shall now turn to an important use of P_5, the postulate of mathematical induction, which was considered to be perhaps the most basic axiom of mathematics in the opinion of Henri Poincaré (1854–1912), one of the greatest of modern mathematicians.

144. Mathematical induction. Notice that P_5 provides us with a criterion for deciding whether a given set S of natural numbers contains *all* natural numbers. If we can somehow prove that the given set S actually has the properties (1) and (2) of P_5, then the axiom asserts that S contains all natural numbers. This criterion may be used to prove theorems like the following. The method of proof is called **mathematical induction.**

Example 1. Prove that if n is any natural number whatever, the formula

$$(A) \qquad 2 + 4 + 6 + 8 + \cdots + 2n = n(n+1)$$

is correct.

First let us see what the theorem asserts. On the left we have the sum of the first n even numbers. The first even number is $2 \cdot 1$; the second even number is $2 \cdot 2$; the third even number is $2 \cdot 3$; the fourth even number is $2 \cdot 4$; the nth even number is $2n$. The theorem then asserts that the sum of the first n even numbers is exactly $n(n+1)$ no matter what natural number n is. For example, it asserts that the sum of the first 50 even numbers is $50 \cdot 51 = 2550$. Hence, the theorem makes an infinite number of assertions, one assertion corresponding to each value of n.

For example, the theorem says that:

for $n = 1$, $\quad 2 = 1(1+1)$;
for $n = 2$, $\quad 2 + 4 = 2(2+1)$;
for $n = 3$, $\quad 2 + 4 + 6 = 3(3+1)$;
for $n = 4$, $\quad 2 + 4 + 6 + 8 = 4(4+1)$;
for $n = 5$, $\quad 2 + 4 + 6 + 8 + 10 = 5(5+1)$;

and so on. The five statements above can be verified directly by arithmetic. But saying "and so on" is begging the question, or simply assuming that things do go on as we wish. We cannot prove that this formula works for *all* natural numbers n by merely going on to verify it for one natural number after another because neither we nor our posterity can live to finish the job. Instead of futilely going on with our direct verification of the formula for particular values of n, we use axiom P_5, as follows.

Let S be the set of all those natural numbers for which the theorem is true. We have already verified by direct arithmetical calculation that 1 is in the set S; this is condition (1) of P_5. That is, the formula is true for the value $n = 1$. Let us now prove that condition (2) of P_5 is satisfied. Condition (2) says that *if* k is a natural number for which (A) is true (that is, if k is in S) *then* $k + 1$ is also a natural number for which the theorem is true (that is, $k + 1$ is in S). We must prove the following theorem.

Hypothesis. $2 + 4 + 6 + \cdots + 2k = k(k+1)$

Conclusion. $2 + 4 + 6 + \cdots + 2k + 2(k+1)$
$= (k+1)(k+1+1)$
$= (k+1)(k+2) = k^2 + 3k + 2.$

The hypothesis is what the formula asserts for the value $n = k$; it is obtained by substituting $n = k$ in the proposed formula (A). The conclusion is what the formula asserts for the value $n = k + 1$; it is obtained by substituting $n = k + 1$ in the proposed formula (A). We have to show that if the hypothesis is a true statement then so must the conclusion be a true statement. That is, it is the formula (A) which is on trial. We have to show that if (A) tells the truth for $n = k$, then it also tells the truth when $n = k + 1$.

Proof. By hypothesis,

$$2 + 4 + 6 + \cdots + 2k = k(k + 1).$$

The left side is the sum of the first k terms of the sequence of even numbers. But the left side of the conclusion is the sum of the first $k + 1$ terms of the sequence of even numbers. It is this latter sum about which we must prove something. Hence, it is natural to add to both sides of our hypothesis the $(k + 1)$th or next term of the sequence; that is, the $(k + 1)$th even number $2(k + 1)$. We may add $2(k + 1)$ to both sides, since if equals are added to equals the results are equal. Thus, we obtain

$$2 + 4 + 6 + \cdots + 2k + 2(k + 1) = k(k + 1) + 2(k + 1) = k^2 + 3k + 2,$$

which is precisely the right side of our conclusion. This is what we had to prove.

Since we have verified conditions (1) and (2) of P_5, it follows that *all* natural numbers are in the set S of natural numbers for which the formula (A) is true. That is, formula (A) is true for all natural numbers n.

The work may be arranged conveniently as in the following example.

Example 2. Prove that $3 + 6 + 9 + 12 + \cdots + 3n = \dfrac{3n(n + 1)}{2}$ for any natural number n.

Proof. Let S be the set of natural numbers for which the theorem is true.

Part I. The set S contains 1, because for $n = 1$, the theorem asserts that $3 = \dfrac{3 \cdot 1(1 + 1)}{2} = 3$, which is true.

Part II. If the set S contains any natural number k, then it also contains $k + 1$.

Hypothesis. $3 + 6 + 9 + 12 + \cdots + 3k = \dfrac{3k(k+1)}{2}$.

Conclusion. $3 + 6 + 9 + 12 + \cdots + 3k + 3(k+1) = \dfrac{3(k+1)(k+1+1)}{2} = \dfrac{3(k+1)(k+2)}{2}$
$= \dfrac{3k^2 + 9k + 6}{2}$.

Proof. By hypothesis, $3 + 6 + 9 + 12 + \cdots + 3k = \dfrac{3k(k+1)}{2}$. We may add the $(k + 1)$th term, $3(k + 1)$, to both sides, since if equals are added to equals the results are equal. Thus we obtain

$$\begin{aligned} &3 + 6 + 9 + 12 + \cdots + 3k + 3(k+1) \\ &= \frac{3k(k+1)}{2} + 3(k+1) \\ &= \frac{3k(k+1) + 2 \cdot 3(k+1)}{2} = \frac{3k^2 + 9k + 6}{2} \end{aligned}$$

which is the same as the right side of our conclusion. This proves the proposition of Part II.

Since we have verified the conditions (1) and (2) of P_5, it follows from P_5 that *all* natural numbers are in the set S of natural numbers for which the theorem is true.

Remark 1. It is *necessary* to verify *both* conditions (1) and (2). From the example of the tin soldiers it is intuitively clear that both conditions are necessary. Knocking down the first soldier will not accomplish the fall of all the soldiers unless they are properly spaced so that the kth soldier knocks over the $(k + 1)$th soldier as he falls. Similarly having them properly spaced does us no good unless the first soldier is actually knocked down. We give two algebraic examples to show that both conditions are necessary.

(*A*). The (false) formula $2 + 4 + 6 + \cdots + 2n = n(n + 1) + 10$ will satisfy condition (2) but not condition (1). For if k were a number for which the formula were true, $k + 1$ would have to be one too. For the proposition of Part II would proceed as follows.

Hypothesis. $2 + 4 + 6 + \cdots + 2k = k(k + 1) + 10.$

Conclusion.

$$\begin{aligned} 2 + 4 + 6 + \cdots + 2k + 2(k + 1) &= (k + 1)(k + 1 + 1) + 10 \\ &= (k + 1)(k + 2) + 10 \\ &= k^2 + 3k + 12. \end{aligned}$$

Proof. If $2 + 4 + 6 + \cdots + 2k = k(k + 1) + 10$ we may add $2(k + 1)$ to both sides obtaining

$$\begin{aligned} 2 + 4 + 6 + \cdots + 2k + 2(k + 1) &= k(k + 1) + 2(k + 1) + 10 \\ &= k^2 + 3k + 12, \end{aligned}$$

which verifies condition (2).

But $n = 1$ does not make the formula true, since for $n = 1$, the formula says that

$$2 = 1(1 + 1) + 10 = 12.$$

Thus condition (1) is not satisfied.

(*B*). The (false) formula $2 + 4 + 6 + \cdots + 2n = n(n + 1) + (n - 1)$ is true for $n = 1$. But it is not true for $n = 2$. If we tried to verify condition (2) it would not work. Let us try it.

Hypothesis.

$$2 + 4 + 6 + \cdots + 2k = k(k + 1) + (k - 1) = k^2 + 2k - 1.$$

Conclusion.

$$\begin{aligned} 2 + 4 + 6 + \cdots + 2(k + 1) &= (k + 1)(k + 1 + 1) + (k + 1) - 1 \\ &= k^2 + 4k + 2. \end{aligned}$$

Proof. By hypothesis, $2 + 4 + 6 + \cdots + 2k = k^2 + 2k - 1.$ Adding $2(k + 1)$ to both sides we get

$$\begin{aligned} &2 + 4 + 6 + \cdots + 2(k + 1) = \\ &\qquad k^2 + 2k - 1 + 2(k + 1) = k^2 + 4k + 1 \end{aligned}$$

which is *not* the same as the right side of the conclusion. Hence Part II of the proof does not go through.

Remark 2. It is important to distinguish between inductive logic as used in the experimental sciences and mathematical induction which belongs to deductive logic. In experimental science we might test a proposed formula like that of example 1, by verifying it directly for $n = 1, 2, 3, \cdots$ and so on up to some arbitrary number. If it worked for the first few numbers we might conclude that it was probably true for all numbers. If it

worked for the first 100 numbers, we would consider it more probably true. If it worked for the first 1000 numbers we would consider its truth for all natural numbers extremely probable. But this would not be the same as asserting that it is certain. In fact, it is easy to write a formula which is true for the first 5 or 100 or 1000 numbers and false thereafter. For example, we have proved that the formula

$$2 + 4 + 6 + \cdots + 2n = n(n+1)$$

is true for all natural numbers n. Now suppose we alter the right member by writing

$$2 + 4 + 6 + \cdots + 2n = n(n+1) + (n-1)(n-2)(n-3)(n-4)(n-5).$$

Then for $n = 1, 2, 3, 4, 5$, this latter formula is correct since the added term will be zero for these values of n. But for $n = 6, 7, 8, \cdots$ and so on the latter formula will be wrong because the added term will no longer be zero and the right member will therefore no longer yield the correct result. For example, for $n = 6$ the right member yields 162 instead of the correct sum 42. In the same way

$$2 + 4 + 6 + \cdots + 2n = n(n+1) + (n-1)(n-2)\cdots(n-1000)$$

would yield the correct result for all values of n from $n = 1$ to $n = 1000$ but will be wrong thereafter. This shows that 1000 or 1,000,000 verifications cannot prove that the formula is true for *all* natural numbers n. Hence the need for the axiom of mathematical induction P_5. A proof by mathematical induction is not an argument of inductive logic, as in experimental science, but is an argument of deductive logic based on axiom P_5.

Remark 3. When we write a sequence of terms like

$$2, 4, 6, \cdots, 2n, \cdots$$

the term $2n$ is called the **general term** or the **nth term** because, for $n = 1$, $2n = 2$; for $n = 2$, $2n = 4$; for $n = 3$, $2n = 6$; and so on. On many psychological "intelligence" tests people are asked to write the next term of a sequence whose first few terms are given. Thus a typical question would be:

"Write the next term of the sequence 2, 4, 6, 8, 10, 12, $\cdots$."

Now, the answer expected is 14 which is obtained by assuming that the general term is $2n$ as is suggested by the first 6 terms which are given. But the general term might just as well have been

$$2n + (n-1)(n-2)(n-3)(n-4)(n-5)(n-6)$$

in which case the first 6 terms would have been exactly the given ones while for $n = 7$ we would have not 14 but $14 + 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 734$. The general term might also have been

$$2n + (n-1)(n-2)(n-3)(n-4)(n-5)(n-6) \cdot f(n)$$

where $f(n)$ is any polynomial, say, in n. A person of exceptional intelligence who was able to perceive that there is no logical reason to suppose that $2n$ is the general term would be at a loss to answer this question on the "intelligence" test, and he might be ranked as a moron, at least as far as this question is concerned. For the question has no logical answer and is really a problem, not in reasoning, but in clairvoyance, for we must guess which of the many possible general terms the examiner is thinking of.*

Remark 4. While the development of the natural numbers on the basis of axioms P_1–P_5 is due to Peano, (late in the 19th century), the idea of proving theorems by mathematical induction can be traced, in some form, back to the 16th century. Note that, by mathematical induction we prove an infinite number of assertions, in a short space.

Remark 5. It is not hard to show that if we verify the truth of any formula for any particular natural number, (for example, 4) and then verify the second condition of our axiom of mathematical induction, then we have proved the truth of the formula for all natural numbers beginning with the particular one started with (in our example, for 4, 5, 6, and so on).

EXERCISES

Write (a) the 5th term; (b) the 75th term; (c) the kth term; (d) the $(k+1)$th term; (e) the $(k+6)$th term, of the sequence whose nth term is:

* E. T. Bell in his *Men of Mathematics* writes that ". . . when Poincaré was acknowledged as the foremost mathematician and leading popularizer of science of his time he submitted to the Binet tests and made such a disgraceful showing that, had he been judged as a child instead of as the famous mathematician he was, he would have been rated—by the tests—as an imbecile."

1. $4n$. **2.** $5n$. **3.** $2n - 1$. **4.** $\frac{1}{n(n+1)}$. **5.** $3n$. **6.** $n(n+1)$.

(a) *In each of the following exercises verify the correctness of the given formula for $n = 1, 2, 3, 4, 5$.* (b) *Does the work in part* (a) *suffice to establish the correctness of the given formula for all natural numbers n? Explain.* (c) *Prove by mathematical induction that the given * formula is correct for all natural numbers n.*

7. $4 + 8 + 12 + 16 + \cdots + 4n = 2n(n+1)$.

8. $5 + 10 + 15 + 20 + \cdots + 5n = \dfrac{5n(n+1)}{2}$.

9. $1 + 2 + 3 + 4 + \cdots + n = \dfrac{n(n+1)}{2}$.

10. $6 + 12 + 18 + 24 + \cdots + 6n = 3n(n+1)$.

11. $1 + 3 + 5 + 7 + \cdots + (2n-1) = n^2$.

12. $\dfrac{1}{1 \cdot 2} + \dfrac{1}{2 \cdot 3} + \dfrac{1}{3 \cdot 4} + \dfrac{1}{4 \cdot 5} + \cdots + \dfrac{1}{n(n+1)} = \dfrac{n}{n+1}$.

13. $1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + 4 \cdot 5 + \cdots + n(n+1) = \dfrac{n(n+1)(n+2)}{3}$.

14. $1^2 + 2^2 + 3^2 + 4^2 + \cdots + n^2 = \dfrac{n(n+1)(2n+1)}{6}$.

15. $1^3 + 2^3 + 3^3 + 4^3 + \cdots + n^3 = \dfrac{n^2(n+1)^2}{4}$.

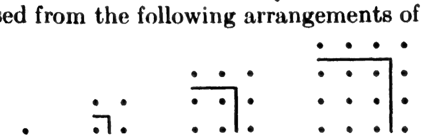
16. $3 + 3^2 + 3^3 + 3^4 + \cdots + 3^n = \dfrac{3^{n+1} - 3}{2}$.

17. Prove by mathematical induction that the sum of the "geometric progression"

$$a + ar + ar^2 + \cdots + ar^{n-1} = \frac{a - ar^n}{1 - r},$$

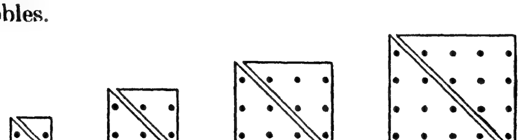
for all natural numbers n.

* Note that while we here prove that the *given* formula is correct we give no hint as to how the right member of the formula was discovered. Some of our examples, however, might have been discovered intuitively. For instance, the formula of exercise 11 might be guessed from the following arrangements of pebbles.



$1 = 1^2$ $\quad 1+3 = 2^2$ $\quad 1+3+5 = 3^2$ $\quad 1+3+5+7 = 4^2$

Similarly the formula of exercise 9 might have been discovered from the following arrangement of pebbles.



$1 = \frac{1 \cdot 2}{2}$ $\quad 1+2 = \frac{2 \cdot 3}{2}$ $\quad 1+2+3 = \frac{3 \cdot 4}{2}$ $\quad 1+2+3+4 = \frac{4 \cdot 5}{2}$

However, the general problem of finding an expression for the sum of n terms of a given sequence is a very difficult one.

18. Prove by mathematical induction that the sum of the "arithmetic progression"

$$a + (a + d) + (a + 2\,d) + \cdots + (a + [n - 1]\,d) = \frac{n(2\,a + [n - 1]\,d)}{2},$$

for all natural numbers n.

19. Prove by mathematical induction that for $n = 4, 5, 6, 7, \cdots$ and so on, $2^n < n(n - 1)(n - 2)(n - 3) \cdots 3 \cdot 2 \cdot 1$. (Hint: see remark 5.)

20. Prove by mathematical induction that $(ab)^n = a^n b^n$ for all natural numbers n.

145. Conclusion. We have been discussing the study of natural numbers on the basis of Peano's postulates. These axioms push the unproved propositions or postulates we start with much further back than the postulates we used in Chapter III. Thus on the basis of Peano's postulates we can prove as theorems the tables of addition and multiplication, and the associative, commutative, and distributive laws, which we took as assumptions in Chapter III. But, in Peano's postulates we still take "natural number" as an undefined term. Hence Peano's abstract mathematical science may be given other concrete interpretations than the one we were thinking of. For example, if the undefined term "natural number" is interpreted to mean "even number," and "successor" to mean "next even number" and "1" to mean the first even number or "2," then all of Peano's postulates are satisfied. Or, if "natural number" is taken to mean "fraction of the form $1/n$" and "1" to mean "1/1" and "successor of $1/k$" to mean "$1/(k + 1)$" then all of Peano's postulates are satisfied. The fact that all arithmetic and algebra can be built up with "natural number" left undefined is not surprising. In fact, a calculating machine is capable of doing arithmetic, but it certainly doesn't know what "number" means. It is, however, constructed so as to obey all the formal rules of numbers, which is all that matters so far as pure mathematics is concerned. In Chapter XV, we shall push the process of definition back further and shall define number in terms of more primitive notions.

REFERENCES

H. Poincaré, *The Foundations of Science*, The Science Press, N. Y., 1913. (Contains the essays, "Science and Hypothesis," "Science and Method," and "The Value of Science.")

B. Russell, *Introduction to Mathematical Philosophy*, Allen and Unwin, London, 1920.

Chapter XV

CARDINAL NUMBERS, FINITE AND TRANSFINITE

146. The definition of cardinal number. Suppose we begin again, as children do, to derive the concept of number from concrete objects. A child first learns to associate the number 3 with a set of three apples, say. After some time it makes a tremendous abstraction in recognizing that a set of three apples, a set of three oranges, a set of three fingers, and the set consisting of Smith, Jones, and Brown all have something in common, namely the property of "threeness." What does it mean to say that these various sets have this property in common? In verifying that the set of three apples has this property in common with the set of three fingers, we pair off each finger with an apple, one for one. This process is called placing the two sets in one-to-one correspondence. In general, two sets of objects are said to be in **one-to-one correspondence** if we have paired off the elements (or members) of one set with the elements of the other set so that to each element of the first set corresponds exactly one element of the second, and to each element of the second there corresponds exactly one element of the first.

For example, the set of all normal living people is in one-to-one correspondence with the set of all living human heads, for to each person corresponds one head and to each head corresponds one person. In monogamous countries the set of all law-abiding husbands is in one-to-one correspondence with the set of all law-abiding wives. The set of all mothers is in one-to-one correspondence with the set of all eldest children. The correspondence between the set of all mothers and the set of all children is not one-to-one but one-to-many.

If two sets A and B can be put into one-to-one correspondence we shall call them **equivalent**; that is, A is equivalent to B or B is equivalent to A. In ordinary language, two equivalent sets *have the same number of elements*. Of course, any set is equivalent to

itself. And it is easy to see that if set A is equivalent to set B and set B is equivalent to set C then set A is equivalent to set C.* Therefore, all the sets equivalent to a given set S are equivalent to each other, or "have the same number" of elements. Consider the class of all sets equivalent to a given set S; for example, the set of all trios or triplets of things. We might conceive of "number" as some property common to all the sets in the class of equivalent sets. Thus "three" might be thought of as some property common to all trios or triplets like the set of letters A, B, C, or the set of people Smith, Jones, and Brown, or the three Muses, or the three little pigs, or the set of wheels on a tricycle, or the set of balls in front of a pawnshop. Now this idea of "three" as an abstract property belonging to all these different sets of objects is somewhat vague. In fact, all these trios might conceivably have more than one property in common; for example, they might all be considered to have the property of "existence." It would be nicer to be able to say exactly what the number "three" is, instead of referring to a vague common property. This can actually be done. There is one simple specific property that every trio has, namely the property of belonging to the class of all sets equivalent to itself. Moreover, no non-trio has this property. Thus we might even more simply say that the number "three" *is* the class of all sets equivalent to the set consisting of Smith, Jones, and Brown. Any set belonging to this class is said to *have* (*or more properly, to belong to*) *the number "three."* Thus we are led to the following general definitions.

Definition 1. *Any two sets are said to* ***have the same number*** *(more properly, belong to the same number) if and only if they are equivalent (that is, if and only if their elements can be placed in one-to-one correspondence.)*

Definition 2. *The* ***number of a set*** *S (or the number of elements in a set S) is the class of all sets equivalent to the set S. A number thus defined is usually called a* ***cardinal number.***†

* In the language of section 19, the relation of equivalence, among sets, is symmetric, reflexive, and transitive.

† If Definition 2 gives you mental indigestion at first, don't let it worry you. The idea of Definition 1 is more important than that of Definition 2 and much easier to digest.

For our own convenience, we invent names for those numbers to which we wish to refer frequently. For example, the cardinal number of any set equivalent to the set consisting of Smith, Jones, and Brown is called "three." The cardinal number of any set equivalent to the set of fingers on one hand is called "five." However, to determine that two sets have the same number, it is not necessary to know what the number of either is called. It suffices to know that the two sets can be placed in one-to-one correspondence. Thus if every seat in a theatre is occupied and there are none standing, we know that the number of seats is the same as the number of people, even if we do not know what this number is called. Our definition 1 corresponds precisely to our intuitive notion of number. In fact, when we count on our fingers we are merely placing some set of objects in one-to-one correspondence with some set of fingers. If the cardinal number is small enough to be in common use we usually invent a name for it to make reference to it easier. Thus the number denoted symbolically by 10^9 or 1,000,000,000 is called a billion. On the other hand the number 10^{100} (or, "one" with a hundred zeros after it) is much larger than anything in common use even in governmental finance. In fact it is larger than the number of electrons in the entire universe, which, according to Eddington, is about 10^{79}. However, in order to refer to this number conveniently, it has been named a "googol," the name having been invented by a child. The much larger number $10^{(10^{100})}$ or 10 raised to the googol power (a "one" with a googol zeros after it) has been called * a "googolplex."

147. Transfinite cardinal numbers. The whole and its parts. A set B is called a **subset** or **part** of the set A if every element (member) of B is also an element of A. According to this definition the set A is a "part" of itself. *If B is part of A and A is also part of B then A and B are the same set.* In other words, if any element of the set A is called an "a" and any element of the set B is called a "b," then the statement "B is part of A" may be worded as "all b's are a's." Similarly, "A is part of B" means "all a's are b's." If all a's are b's *and* all b's are a's then the sets A and B are identical. However, if B is part of A and there exists at least one element of A which is not an element of B then B is

* See Kasner and Newman, *Mathematics and the Imagination.*

called a **proper part** or a **proper subset** of A. That is, a proper part is a part which is not the whole thing.

Consider the set A of *all* natural numbers $1, 2, 3, \cdots, n, \cdots$, and the set B of all natural numbers greater than 10. The set B is a proper part of A since the numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 are in A but not in B. Nevertheless we can place the sets A and B in one-to-one correspondence, by pairing off their elements as follows: let any natural number x in A correspond to the number $x + 10$ in B. Thus,

$$\begin{array}{cccccccc} 1, & 2, & 3, & 4, & \cdots\cdots\cdots, & n & , & \cdots \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & & \updownarrow & & \\ 11, & 12, & 13, & 14, & \cdots\cdots\cdots, & n+10, & & \cdots. \end{array}$$

The double arrow is read "corresponds to." Hence the sets A and B are equivalent, or have the same cardinal number.

Consider the set C of all even natural numbers $2, 4, 6, \cdots, 2\,n, \cdots$. The set C is equivalent to A since they can be placed in one-to-one correspondence by allowing any natural number x in A to correspond to $2\,x$ in C, as follows:

$$\begin{array}{ccccccc} 1, & 2, & 3, & 4, & \cdots\cdots\cdots, & n, & \cdots \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & & \updownarrow & \\ 2, & 4, & 6, & 8, & \cdots\cdots\cdots, & 2\,n, & \cdots. \end{array}$$

Hence C has the same cardinal number as A, in spite of the fact that the set C is a proper part of the set A.

Consider the set D of all odd natural numbers $1, 3, 5, \cdots, 2\,n - 1, \cdots$. The set D is equivalent to A since they can be placed in one-to-one correspondence by allowing the natural number x in A to correspond to the odd number $2\,x - 1$ in D, as follows:

$$\begin{array}{ccccccc} 1, & 2, & 3, & 4, & \cdots\cdots\cdots, & n & , \cdots \\ \updownarrow & \updownarrow & \updownarrow & \updownarrow & & \updownarrow & \\ 1, & 3, & 5, & 7, & \cdots\cdots\cdots, & 2\,n-1, & \cdots. \end{array}$$

Hence D has the same cardinal number as A, although the set D is a proper part of the set A.

We have shown that the set A has the same number of elements (or is equivalent to, or has the same cardinal number as) the sets B, C, and D, each of which is a proper part of A. This shows that the number of elements in a proper part of a set may be quite the same as the number of elements in the whole set.

This is at first surprising since it seems to upset violently the notion (once considered self-evident) that the whole is greater than any of its parts.

Another illustration is the following. Consider a 12 inch line segment AB and a 6 inch line segment CD placed as in Fig. 216.

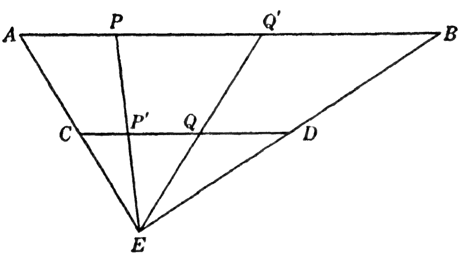


FIG. 216

Let E be the point where AC and BD intersect. Then let P be any point on AB. The line PE intersects CD in exactly one point P'. Conversely, if Q is any point on CD, the line QE intersects AB in exactly one point Q'. Thus the points of AB can be placed in one-to-one correspondence with the points of CD simply by letting any point P on AB correspond to the point P' on CD where the line PE meets CD. Hence the number of points on the 12 inch line segment AB is exactly the same as the number of points on the 6 inch line segment CD. On the other hand the number of points on CD is the same as the number of points on half of AB, as can be seen by setting up a correspondence between AM (M being the midpoint of AB) and CD by means of perpendiculars as in Fig. 217. Since the number of points on AB is the same as the number of points on CD and the number of points on CD is the same as the number of points on half of AB, it follows that the number of points on AB is the same as the number of points on half of AB. In the same way it can be shown that the number of points on any line-segment is the same as the number of points on any other line-segment, regardless of their lengths. There are just as many points on a line-segment one inch long as on one stretching from here to the star Arcturus.

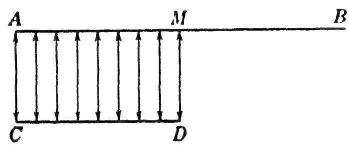


FIG. 217

This illustration also seems to do violence to the notion that the whole is greater than any of its parts. Note, however, that we are talking about the number of points in the line-segment, not its length. It is true that the cardinal number of a set may

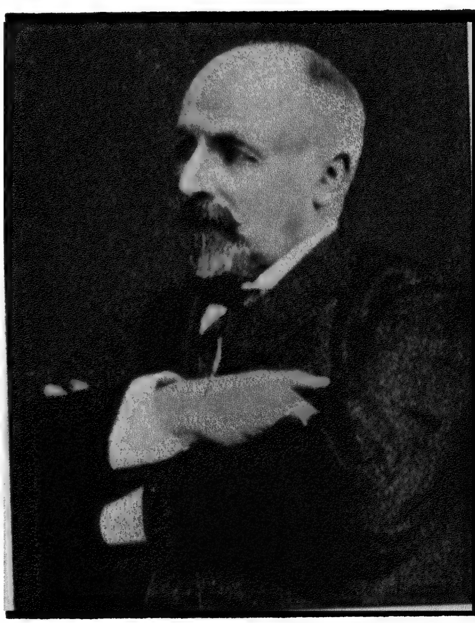
be the same as the cardinal number of a proper subset or part. However, this cannot happen with finite sets, that is, intuitively speaking, with sets that can be counted out at a steady rate of speed in a limited time. So far as finite quantities, or lengths, areas, angles, and other magnitudes used in the elementary applications of mathematics are concerned, "the whole is greater than any of its (proper) parts" remains valid. In fact, it is customary to *define* an infinite set as follows.

DEFINITION 1. *A set is* ***infinite*** *if it is equivalent to a proper part of itself. If a set is not infinite it is called* ***finite.***

DEFINITION 2. *The cardinal number of an infinite set is called a* ***transfinite*** *cardinal number. The cardinal number of a finite set is called a* ***finite*** *cardinal number.*

Remark. To say that the number of elements in a set is infinite is quite different from saying that is is a very large finite number. Thus 10^{100} and $10^{(10^{100})}$ are staggeringly large finite numbers but they are finite. Neither should the idea of infinite or transfinite numbers discussed here be confused with the idea of "infinity" which we discussed in connection with sequences, limits, and functions in section 103.

The cardinal number of any set equivalent to the set of all natural numbers is denoted by $\aleph_0$ (read, "aleph-null") after G. Cantor (1845–1918) who first studied transfinite numbers systematically. The cardinal number of the set of all points on the line-segment AB is denoted by $\aleph$ (read, "aleph"). Both are transfinite cardinal numbers. It can be shown that $\aleph_0$ is not the same as $\aleph$. In fact, there are many differ-



Georg Cantor
1845–1918, German

ent transfinite cardinals which can be studied systematically.

An arithmetic of transfinite numbers can be worked out. For example, we shall define the "sum" of two cardinal numbers, finite or transfinite, as follows. Let x be the cardinal number of a set X and let y be the cardinal number of a set Y.

DEFINITION 3. *If the sets X and Y have no element in common, then by the* **sum** $x + y$ *of the two cardinal numbers x and y we shall mean the cardinal number of the set consisting of all the elements of the set X and all the elements of the set Y together.*

In other words, the sum of the cardinal numbers of two sets is defined to be the cardinal number of the set composed of all the elements of the two sets taken together, *provided the two given sets have no element in common.* The reason for this latter restrictive proviso becomes obvious when you consider the following simple example. Let X be the set consisting of Ames and Brown and let Y be the set consisting of Camp, Davis and Earl. Then the sum of the cardinal number of set X (2) and the cardinal number of the set Y (3) is the cardinal number of the set consisting of Ames, Brown, Camp, Davis, and Earl together (5). But if the set Y were composed of Brown, Camp and Davis we would not say that the sum of 2 and 3 is the cardinal number of the set composed of Ames, Brown, Camp, and Davis.

Let us apply this natural definition of the sum of two cardinal numbers to transfinite numbers.

Example 1. The set A of all natural numbers has the cardinal number $\aleph_0$. As we showed above, the set C of all even natural numbers has the same cardinal number $\aleph_0$ and the set D of all odd natural numbers also has the cardinal number $\aleph_0$. But the set A of all natural numbers has for its elements all the elements of the sets C and D together; and the sets C and D clearly have no element in common. Hence the sum of the cardinal number of C plus the cardinal number of D is the cardinal number of A, or

$$\aleph_0 + \aleph_0 = \aleph_0.$$

Example 2. Let E be the set of numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. The cardinal number of the set E is 10. The set B of all natural numbers greater than 10 has the cardinal number $\aleph_0$ as

shown above. But all the elements of the sets B and E together constitute the set A of all natural numbers; and the sets E and B have no element in common. Hence

$$\aleph_0 + 10 = \aleph_0.$$

These are only two of the many surprising statements that arise in the arithmetic of transfinite numbers.

Another surprising result is the fact that there are just as many natural numbers as there are positive fractions, despite the fact that the positive fractions are densely distributed on the line while the natural numbers are spaced apart at intervals of one unit. (See section 24.) This may be seen by arranging the fractions in the order indicated by the arrow in Fig. 218 as follows:

$$\begin{array}{cccccccccccc}
\frac{1}{1}, & \frac{1}{2}, & \frac{2}{1}, & \frac{3}{1}, & \frac{2}{2}, & \frac{1}{3}, & \frac{1}{4}, & \frac{2}{3}, & \frac{3}{2}, & \frac{4}{1}, & \frac{5}{1}, & \cdots \\
\updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \\
1, & 2, & 3, & 4, & 5, & 6, & 7, & 8, & 9, & 10, & 11, & \cdots.
\end{array}$$

In this way we are able to place the set of all positive fractions in one-to-one correspondence with the set of all natural numbers. Hence the cardinal number of the set of all positive fractions is $\aleph_0$, just as is the cardinal number of the set of all natural numbers. Of course, the set of all natural numbers is a proper subset of the set of all positive fractions; they occupy only the first horizontal row in Fig. 218.

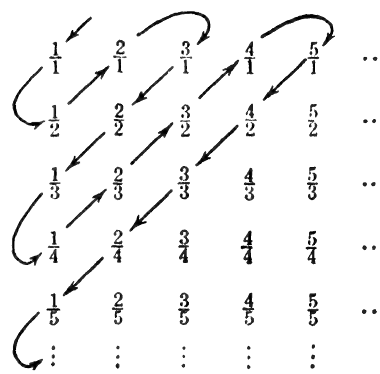


FIG. 218

The subject of infinity has been difficult since ancient times. Cantor made the first extensive assault upon it that could be called successful, although others (Galileo, for example) had had glimmerings of his ideas before. Cantor was led to the study of transfinite numbers by his researches in connection with functions and real numbers. His work provided a new foundation for much of mathematics and stirred up many controversies which are not yet settled.

EXERCISES

1. Show that the set of all square integers 1, 4, 9, 16, 25, $\cdots$, n^2, $\cdots$ has the cardinal number $\aleph_0$.

2. Show that the set of all positive fractions with numerator 1 has the cardinal number $\aleph_0$.

3. Show that the set of points on the hypotenuse of a right triangle is equivalent to the set of points on the shortest leg of the triangle.

4. Show that $\aleph_0 + 3 = \aleph_0$.

5. Show that $\aleph_0 + 7 = \aleph_0$.

148. The finite cardinal numbers. We pointed out in Chapter XIV that, in Peano's postulates for the natural numbers, the term "natural number" was an undefined term. Suppose we now interpret the term "natural number" as meaning "finite cardinal number." Let us consider a set consisting of a single object. Its cardinal number may be called 1. In Peano's system the term "successor" was undefined. Let us now interpret the term "successor" as follows. Consider any finite set A of objects. Let the cardinal number of A be denoted by a. Let us consider a set B consisting of a single object *not* belonging to A. Now consider the set consisting of all the elements of A together with the element of B. The cardinal number of this set is called the successor of a or $a + 1$. With these concrete interpretations for the undefined terms in Peano's abstract system, we could now prove that all his postulates are satisfied. This will not be done here. While Peano's postulates describe abstractly "the process of counting," the finite cardinal numbers provide a concrete interpretation of Peano's abstract mathematical science which really has to do with the "number of objects in a set" as we think of this phrase intuitively. Our definition of cardinal number involves the concepts of "set" of objects, and "one-to-one correspondence," both of which are very primitive intuitive ideas and are left undefined here.

REFERENCES

E. T. Bell, *The Queen of the Sciences*, G. E. Stechert, N. Y., 1938.

G. Cantor, *Contributions to the Founding of the Theory of Transfinite Numbers* (tr. by Jourdain), Open Court, Chicago, 1915.

T. Dantzig, *Number, the Language of Science*, Macmillan, N. Y., 1939.

Galileo Galilei, *Two New Sciences*, Macmillan, N. Y., 1914.

E. Kasner and J. Newman, *Mathematics and the Imagination*, Simon & Schuster, N. Y., 1940.

B. Russell, *Introduction to Mathematical Philosophy*, Allen & Unwin, London, 1920.

J. W. Young, *Fundamental Concepts of Algebra and Geometry*, Macmillan, N. Y., 1911.

Chapter XVI

EUCLIDEAN AND NON-EUCLIDEAN GEOMETRY

149. Introduction. In Chapter II, section 8, we pointed out that Euclidean geometry, considered as a branch of pure mathematics was an abstract mathematical science. We took as undefined terms the words point and line (and others); and we assumed many unproved propositions or postulates such as the following.

POSTULATE 1. *Given any two distinct points, there is at least one line containing them.*

POSTULATE 2. *Given any two distinct points, there is at most one line containing them.*

These two postulates are sometimes quoted together as "two points determine exactly one line." From these two postulates alone we can deduce a theorem.

THEOREM 1. *Two distinct lines have at most one point in common.*

Proof. Let l and m be any two distinct lines. Either l and m have more than one point in common or they have not. Suppose they have more than one point in common; that is suppose the theorem were false. Then there would be two distinct points, P and Q, say, each of which is on both lines. Then by Postulate 2, l and m must be the same line, since there can be at most one line through P and Q. This contradicts the hypothesis that they are distinct. Since our supposition leads to a false conclusion, it must be false. Therefore, the only remaining possibility is that l and m do not have more than one point in common. This is what we had to prove.

This theorem tells us that any two distinct lines have either no point or one point in common. We now make the following definition.

DEFINITION. *If two distinct lines in the same plane have no point in common they are called* ***parallel*** *to each other.*

In other words, two lines are called parallel if they do not meet.*

Notice that Theorem 1 does not tell us whether there *are* any parallel lines at all. In Euclidean geometry we proceed to make the following **Euclidean parallel postulate.**

POSTULATE 3. *Given a line l and a point P not on l, there exists one and only one line m through P which is parallel to l.*

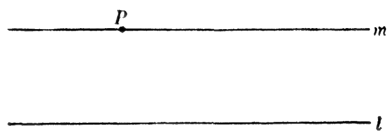


FIG. 219

Many other postulates are made in Euclidean geometry besides these three but we shall not trouble to list them here. The important thing is that, having given a complete list of postulates, it is possible to deduce all the theorems by pure logic without ascribing any meaning to the undefined terms like "point" and "line" just as was done with the simple arguments in section 5, Chapter II. In fact, "point" and "line" might equally well be called "mumbo" and "jumbo" respectively to emphasize that they may be regarded as completely undefined terms. As we saw in section 87, these undefined terms are susceptible of more than one concrete interpretation; thus "point" may be interpreted either as a dot or as a pair of real numbers. But in deducing theorems we need not have any concrete interpretation in mind. In particular, no reference to diagrams need be made.

Exercise. Rewrite Postulates 1 and 2, and Theorem 1 and its proof, using mumbo and jumbo instead of the words point and line, respectively.

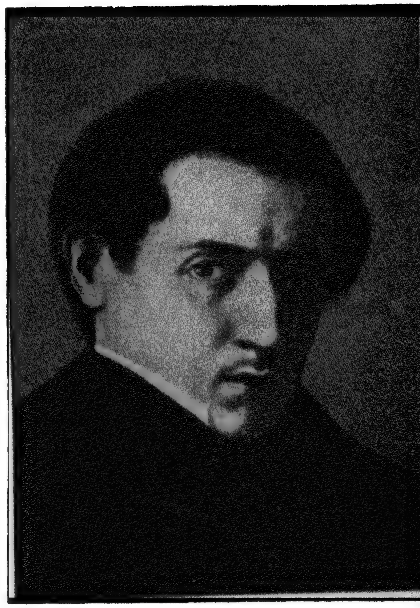
If one thinks of Euclidean plane geometry as referring to diagrams, then one is dealing with Euclidean geometry not as an abstract mathematical science, but as a concrete interpretation or application to the real world where the undefined terms point and line are interpreted as referring to the dots and streaks drawn on paper. We have already remarked, in section 8, Chapter II,

* If follows from this definition that any statement about parallel lines "meeting" at "infinity" or at any other forsaken place is sheer nonsense of the most self-contradictory sort, since it would mean that "lines which never meet, meet at infinity." Despite the fact that such statements are as silly as the statement "two railroads which never meet, meet in Chicago," they are often made.

that when we make this concrete interpretation of Euclidean geometry we can no longer be sure that our postulates are true statements concerning these objects. The idea that these postulates are self-evident truths derived from our physical experience does not hold water. If the undefined terms point and line are taken to mean dot and streak respectively then even postulate 2 is no longer true except in an approximate sense, as may be seen by looking at a diagram with a magnifying glass (Fig. 220), although intuitively it seems as though there would be only one streak through two dots if you could only make the dots and streaks thin enough (you can't). In fact, the Greeks probably thought of points and lines as "idealized" dots and streaks, whatever they may be. Thus Euclid and nearly everybody else for 2000 years after him probably thought of these postulates as self-evident truths about these "idealized" points and lines and they regarded Euclidean geometry as an absolutely true description of physical space. Fortunately, however, many people regarded the parallel postulate as being somewhat less self-evident than the others. This is natural because the parallel postulate asserts that certain lines will not meet no matter how far they are produced. Now, if you wish to regard your axioms as self-evident truths derived from our physical experience, you can hardly say that you have any experience about lines which are produced indefinitely far. Our experience is confined to a very small portion of what we imagine to be the real world. Therefore, many people did not like to regard the parallel postulate as self-evident even though they were content to regard the other postulates as self-evident. There is some evidence that Euclid himself was reluctant to assume the parallel postulate and postponed its use as long as he could, using proofs which did not involve it as far as he could. Since the parallel postulate was not regarded as a self-evident truth in good standing, it was natural to try to prove it as a theorem on the basis of the other postulates. If it could be proved as a theorem, there would be no need to assume it and the whole question of its self-evidence would disappear, since it would then be true automatically by virtue of being a logical consequence of the other self-evidently true postulates. Therefore,

FIG. 220

for about 2000 years, attempts were made to prove the parallel postulate on the basis of the other postulates, but without success. The history of these attempts is interesting but can hardly be gone into here. The question was definitely settled in the 19th century when Lobachevski (1793–1856), a Russian, J. Bolyai (1802–1860), a Hungarian, and Gauss (1777–1855), independently of each other, established definitely that the parallel postulate could not be proved on the basis of the other postulates. That is, they showed that it is entirely independent of the other postulates.



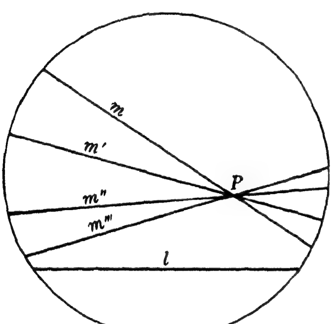
Nicholaus Ivanovich Lobachevski
1793–1856, Russian

How they showed this is an interesting story. Essentially their method was this. If the Euclidean parallel postulate can be proved on the strength of the other postulates, then the other postulates together with a postulate that flatly contradicts the Euclidean parallel postulate must lead ultimately to a contradiction. To be precise, suppose all of Euclidean geometry can be deduced from 16 postulates, of which the 16th is the Euclidean parallel postulate. Now if the 16th statement can actually be deduced from the first 15 then a new logical system based on the first 15 and a new 16th which contradicts the Euclidean parallel postulate must be inconsistent, that is, must lead to a contradiction. For the new system would contain among its theorems the Euclidean parallel postulate, as a consequence of the first 15 assumptions, and this would contradict our new 16th postulate. That is, an abstract mathematical science based on all the other postulates of Euclidean geometry together with a postulate which contradicts the Euclidean parallel postulate would have to be inconsistent. What Lobachevski did was to assume, *instead* of Euclid's parallel postulate, the following postulate.

Lobachevskian Parallel Postulate. *Given a line l and a point P not on l, there exist at least two lines through P parallel to l.*

This postulate was assumed together with *all* the *other* postulates of Euclid. The resulting abstract mathematical science is usually called **Lobachevskian plane geometry** (or hyperbolic geometry) because Lobachevski was the one who developed this geometry to the greatest extent. It is one of the many forms of **non-Euclidean geometry.** As remarked above, Lobachevski established the independence of the parallel postulate by showing that this Lobachevskian geometry is not inconsistent, that is, does not lead to a contradiction. How this is done will now be described. We confine ourselves to plane geometry.

150. A Euclidean model for Lobachevskian geometry. We shall show that the abstract mathematical science called Lobachevskian geometry can be given a concrete interpretation in a part of the Euclidean plane. Take a definite circle, as small or as large as you please, in the Euclidean plane. By the *Lobachevskian plane* we shall mean the interior of this circle. Let us now interpret the undefined term "point" in Lobachevski's system as meaning a Euclidean point interior to this circle. Let us interpret the undefined term "line" in Lobachevski's system as meaning that part of a Euclidean line which is contained within the circle. Then, it is easy to verify that, when the undefined terms point and line are given these concrete interpretations, all of Lobachevski's postulates are satisfied, including *his* parallel postulate (Fig. 221). This shows that his abstract mathematical science is consistent; more precisely, it shows that if Lobachevski's geometry involves any inconsistencies or contradictions then so does Euclid's, since we have a model or concrete interpretation of the Lobachevskian "plane" within the Euclidean "plane." That is, any contradictory statements in Lobachevskian geometry would yield corresponding contradic-



None of the "lines" m, m', m'', m''' have any "points" in common with the "line" l; that is, they are all "parallel" to l.

FIG. 221

tory statements concerning figures within a circle in Euclidean geometry. What is really established is that *Lobachevskian geometry is as consistent as Euclidean.* As far as logical consistency is concerned, there is no reason to prefer Euclid's geometry to Lobachevski's. The question of how *any* abstract mathematical science can be proven absolutely consistent is a very deep problem; it is obviously not easy to prove that no contradiction will *ever* appear among the logical consequences of a set of postulates no matter how many theorems are deduced. We return to this question in Chapter XVII. In any case, if we grant that Euclidean geometry is consistent, then so is Lobachevskian.

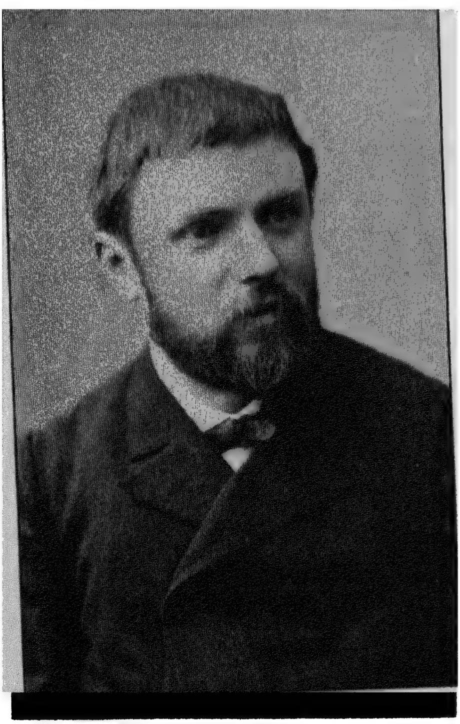
However, the reader may now object that while Lobachevskian geometry may be a logically self-consistent system, it is not at all clear that it lends itself to practical application to the real world. Let us examine this question.

Since, Lobachevski's geometry differs from Euclid's only with respect to the parallel postulate, all theorems whose proofs do not depend on the parallel postulate will be quite the same in both geometries. But we should expect theorems which depend on the parallel postulate to differ. For example, in Euclidean geometry we can prove that the sum of the angles of a triangle is exactly 180°, while in Lobachevski's geometry we can prove that the sum of the angles of a triangle is less than 180°. (How this can be proved will be indicated in section 151.) The reader may now ask: "Does not this allow us to say that Lobachevskian geometry does not apply to the real world? For if we measure the angles of a triangle their sum will be 180°." To dispose of this question we have only to remind the reader that all physical measurements are at best approximate and that we could not tell by measurement whether the sum of the angles of the triangle is 180° or 179.99999999876° (or, for that matter, 180.0000002°). Thus this test does not enable us to say that Lobachevskian geometry does not apply to the real world.

It can also be proved that, in Lobachevskian geometry, the sum of the angles of a triangle depends on the area of the triangle, differing from 180° more and more as the area becomes larger. But even this does not enable us to assert that Lobachevskian geometry in inapplicable to the real world. For, small and large have only comparative, not absolute meaning. For ex-

ample, 1,000,000 is large compared to .0001 but 1,000,000 is small compared to 10^{100}; and .0001 is large compared to .00000000001. If we measured what we felt was a very large triangle, like a triangle joining three distant stars, and found the sum of the angles to be indistinguishably close to 180°, we could not conclude that Lobachevskian geometry does not provide an accurate description of the real world. For it might well be that our complete observational experience with the real world, large though it may seem to us, is confined to a region so small that within it the discrepancies between Lobachevskian and Euclidean geometry are still too small to measure.

It must not be supposed that the interior of a circle is the only possible concrete interpretation of Lobachevskian geometry. The whole point is that the actual universe, so far as we can tell, may be a concrete interpretation of Lobachevskian geometry. The interior of the circle was introduced merely to establish that Lobachevskian geometry is just as logically consistent as Euclidean since concrete objects in Euclidean geometry satisfy the Lobachevskian postulates. In particular, it must not be inferred from the circle-interior model of Lobachevskian geometry that the length of a line is limited in Lobachevskian geometry. This model is only one possible picture or map of Lobachevskian geometry and one must not read from it more than is intended; just as one must not conclude from a map of the United States that the state of Kansas has an area of less than 3 square inches and is uniformly pink in color. In fact, some geographical maps (Mercator maps, in which meridians of longitude appear as



Jules Henri Poincaré

1854–1912, French

vertical straight lines and parallels of latitude as horizontal straight lines) do not even preserve the relative proportions of distances or areas.

On the other hand, the entire universe might be contained within a large sphere, so that any physical "plane" would be actually contained within a large circle, whose radius may be supposed to be larger than the range of our actual experience. Of course we are used to imagining the universe as extending indefinitely but this is merely a habit of thought and not a logical necessity. In fact, Poincaré suggested the possibility that we might be convinced that the world were unlimited in extent even though it were really limited. His argument is essentially this. Suppose the world were contained in a limited sphere, but that we ourselves, and all other material bodies, shrank as we approached the boundary in such a way that we could never reach it. Then we would think that the world were limitless in extent. Furthermore, we could never detect our own shrinkage because our measuring rods would shrink in proportion.

To summarize, the discrepancies between the theorems of Euclidean and Lobachevskian geometry may be smaller than we are able to measure, so that for practical applications there is no choice between them. In practical work we prefer Euclidean geometry because it is more familiar and because it is easier to handle.

151. Proofs of some theorems in Lobachevskian geometry. We shall sketch a proof of the theorem that, in Lobachevskian geometry, the sum of the angles of any triangle is less than 180°. This will be a fairly easy (though long) job to do because Lobachevskian geometry is very much like Euclidean, since it differs *only* in so far as the parallel postulate is concerned. Therefore all the theorems of Euclidean geometry whose proofs do not involve the parallel postulate are also correct theorems in Lobachevskian geometry. Hence, we shall not bother to prove many such theorems here, but we shall assume you are familiar with them from your study of Euclidean geometry in high school. *Until we say otherwise explicitly, all the following theorems are based only on those postulates which are common to Euclidean and Lobachevskian geometry; that is, they involve neither the Euclidean*

nor the Lobachevskian parallel postulate, and therefore are correct theorems in both geometries. Such propositions as "vertical angles are equal," and others, will be used without proving them here.

Remark. It will be necessary to recall that the statement

(1) "if A is true then B is true"

does not imply the statement

(2) "If A is not true, then B is not true"

nor does (1) imply the statement

(3) "If B is true then A is true."

(3) is called the converse of (1). (2) and (3) are equivalent. (2) and (3) may be incorrect even though (1) is correct. Likewise (2) and (3) may be correct even though (1) is not.

Exercise. Make up a simple example to illustrate each of these points.

We shall not prove Theorems I and II.

THEOREM I. *Two triangles are congruent if two sides and the included angle of one are equal respectively to two sides and the included angle of the other.*

COROLLARY Ia. *The base angles of an isosceles triangle are equal.*

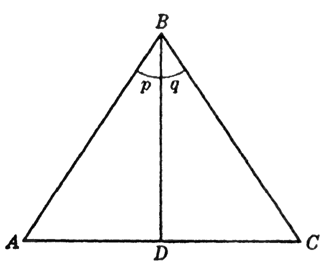


FIG. 222

Proof. Draw the angle bisector BD of the vertex angle (Fig. 222). Then $AB = BC$ by hypothesis, $p = q$ and $BD = BD$. Therefore, triangles ABD and BCD are congruent, by Theorem I. Hence $A = C$, since corresponding parts of congruent triangles are equal.

THEOREM II. *At least one perpendicular can be drawn to a given line l from a given point P, whether P is on l or not.*

We can now prove the following theorems.

THEOREM III. *If the point P is on the line l, only one perpendicular can be drawn to l at P.*

Proof. If there were two, then angle a and angle b (Fig. 223) would be right angles. But $a > b$, since the whole is greater than

any of its parts, which is a contradiction since all right angles are equal. Hence there can be only one perpendicular at P.

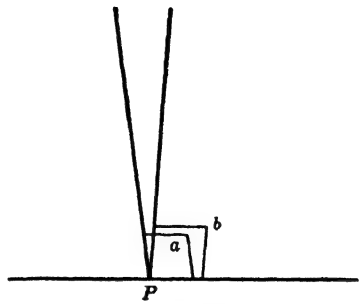


FIG. 223

THEOREM IV. *An exterior angle of a triangle ABC is greater than either remote interior angle.*

Proof. We prove first that the exterior angle DCB is greater than angle B (Fig. 224). Let M be the midpoint of BC. Thus $BM = MC$. Extend AM its own length to E, so that $AM = ME$. Angle q = angle r, since vertical angles are equal. Therefore triangles ABM and CEM are congruent by Theorem I. Hence angle B = angle s because corresponding parts of congruent triangles are equal. Now, the exterior angle p ($\measuredangle DCB$) is greater than angle s, since the

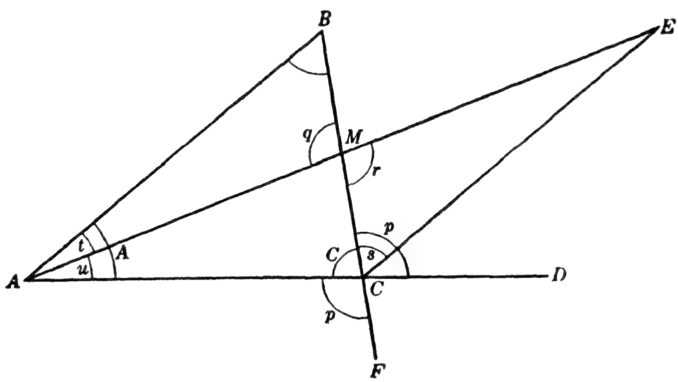


FIG. 224

whole is greater than any of its parts. Hence angle DCB is greater than angle B. To prove that $\measuredangle DCB$ or $\measuredangle ACF$ is greater than $\measuredangle A$, we would start by taking the midpoint of AC and proceed in the same way.

Exercise. Prove the last statement in detail.

COROLLARY IVa. *If two alternate interior angles made by a transversal to two lines are equal, then the lines are parallel.*

Proof. Let angle p = angle q (Fig. 225). Either the lines are parallel or not. Suppose the lines were not parallel. Then they

would meet at some point C (Fig. 225). Then $p > q$ by Theorem IV, contradicting the hypothesis that $p = q$. Since our supposition leads to a false conclusion, it is false. Hence the lines are parallel.

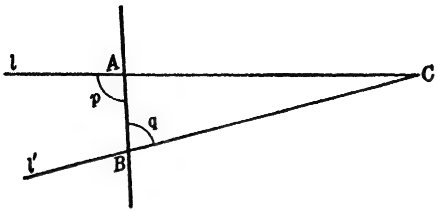


Fig. 225

Remark. The converse of IVa cannot be proved without the Euclidean parallel postulate; it is not true in Lobachevskian geometry.

Theorem V. *The sum of any* **two** *angles of a triangle ABC is less than 180°.*

Proof. Consider the exterior angle p ($\measuredangle BCD$) at C (Fig. 224). By Theorem IV, $p > B$ and $p > A$. Now $p + C = 180°$ since their sum is a straight angle. Hence $B + C < 180°$ and $A + C < 180°$. To prove $A + B < 180°$, use an exterior angle at A and proceed similarly.

Exercise. Prove the last statement in detail.

By the **angle-sum of a triangle** we mean the sum of its three interior angles. (We must not assume that this angle-sum is 180° because the proof of that theorem in Euclidean geometry depends on the parallel postulate which we are not assuming here.)

Theorem VI. *Let A be any angle of a triangle ABC. Then there is another triangle one of whose angles is A/2 and whose angle-sum is equal to the angle-sum of triangle ABC.*

Proof. Suppose the given triangle ABC is that of Fig. 224. As in the proof of Theorem IV, we have triangle ABM congruent to triangle ECM. Then $t = E$, $B = s$, $C = C$. Hence $E + u + (C + s) = (t + u) + C + B$, by substitution, or $E + u + \measuredangle ACE = A + C + B$. Thus the triangle ACE has the same angle-sum as triangle ABC. Now since $t + u = A$ either t or u must be $\leqq A/2$, since if both were $> A/2$ their sum would be $> A$. But $t = E$. Hence either E or u must be $\leqq A/2$. Thus triangle ACE satisfies the requirements of the theorem. This completes the proof.

Remark. We must now point out the arithmetic fact that if A is any positive real number, no matter how large, and h is any positive number, no matter how small, then there exists a natural number n such that $A/2^n < h$. That is, no matter how small h is, there exists a definite term in the sequence $A/2$, $A/4$, $A/8$, $A/16$, $\cdots$, $A/2^n$, $\cdots$ which is smaller than h. For example, suppose $A = 200$ and h is $1/10$. Then the sequence is 100, 50, 25, 12.5, 6.75, 3.375, 1.6875, .84375, .421875, .2109375, .10546875, .052734375, $\cdots$. The last term we wrote is less than $1/10$. Clearly if we divide any given number by 2 often enough, we can make the result as small as we please.

THEOREM VII. *The angle-sum of any triangle is not greater than 180°. That is, if ABC is any triangle, then* $A + B + C \leqq 180°$.

Proof. Either the theorem is correct or not. Suppose the theorem were not correct. Then we would have $A + B + C > 180°$, or $A + B + C = 180° + h°$ where h is some positive number; for example, we might have $A + B + C = 180° + 1°$ or $181°$. By Theorem VI, there is another triangle T_1 one of whose angles A_1 is $\leqq A/2$ and which has the same angle-sum as triangle ABC. Applying Theorem VI now to triangle T_1, there is another triangle T_2 one of whose angles A_2 is $\leqq A_1/2 \leqq A/2^2$ and which has the same angle-sum as triangle ABC. This argument can be repeated as often as we please. After n such steps we obtain a triangle T_n one of whose angles $A_n \leqq A/2^n$ and which has the same angle-sum as triangle ABC. Now, by the remark above, there exists a natural number n, such that $A/2^n < h$. For such a value of n, one of the angles in triangle T_n is less than h. But the sum of the other two angles of the triangle T_n must be less than 180° by Theorem V. Hence the angle-sum of triangle T_n must be less than $180° + h°$ contradicting the supposition that it is equal to $180° + h°$. Therefore our supposition is false. Hence the theorem is correct. This completes the proof.

THEOREM VIII. *An exterior angle of a triangle is greater than or equal to the sum of the remote interior angles.*

Proof. By Theorem VII, $A + B + C \leqq 180°$. Subtracting C from both sides we obtain $A + B \leqq 180° - C$. But (Fig.

226) $p = 180° - C$. Hence by substitution, $A + B \leqq p$, or $p \geqq A + B$ which was to be proved.

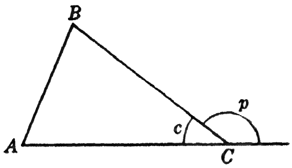


FIG. 226

THEOREM IX. *If the point P is not on the line l, then only one perpendicular can be drawn from P to l.* (Compare Theorem III.)

Proof. If there were two perpendiculars then the angle-sum of triangle PAB (Fig. 227) would be more than 180° contrary to Theorem VII.

DEFINITION. *By the* **angle-sum** *of a quadrilateral is meant the sum of its four interior angles.*

THEOREM X. *The angle-sum of any quadrilateral is $\leqq 360°$.*

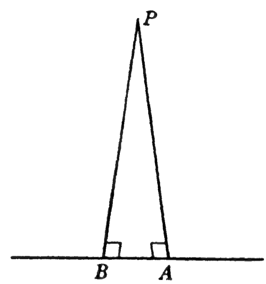


FIG. 227

Proof. Let $ABCD$ be any quadrilateral. Draw a diagonal lying within the quadrilateral, say AC (Fig. 228). Now $p + B + r \leqq 180°$ and $q + D + s \leqq 180°$ by Theorem VII. Hence $(p + q) + B + (r + s) + D \leqq 360°$ or $A + B + C + D \leqq 360°$. This completes the proof.

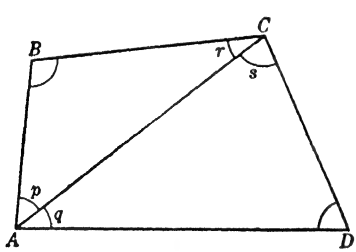


FIG. 228

Remark. Theorem VII does not tell us whether there *are* any triangles whose angle sum is 180° or not. The next few theorems will be based on the hypothesis that there is (at least) *one* such triangle. Of course, there may not really be any, as far as we know at this moment. That does not stop us from investigating what *would* take place *if* there *were* one.

THEOREM XI. *If there exists* **one** *triangle whose angle-sum is 180° then there exists a right triangle whose angle-sum is 180°.*

Proof. Suppose ABC is the triangle with angle-sum 180°, whose existence is assumed in our hypothesis. Then (Fig.

229) we can draw the perpendicular from some vertex, say B, to the opposite side. Now $p + q = B$. Hence by hypothesis, $A + C + (p + q) = 180°$. Now $r = s = 90°$. Hence we have $A + C + (p + q) + r + s = 360°$; or $(A + p + r) + (q + s + C) = 360°$. Now, if one of these parentheses were less than half of 360° the other would have to be more than half of 360°. But each parenthesis is the angle-sum of one of the triangles ABD or BCD and therefore cannot be more than 180° by Theorem VII. Hence, neither of these triangles can have an angle-sum less than 180°. But the angle-sum of any triangle is either less than or equal to 180°. Hence ABD and BCD are right triangles whose angle-sums = 180°. This completes the proof.

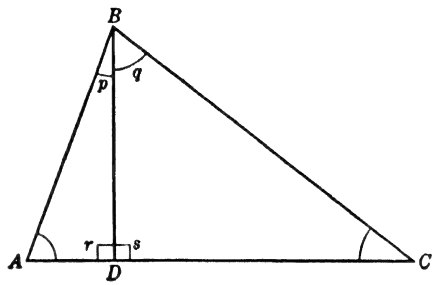


FIG. 229

DEFINITION. *A quadrilateral with four right angles is called a* ***quadrirectangle.***

Notice that if three of the angles of the quadrilateral are right angles, we may *not* conclude that the fourth is a right angle, since the sum of the angles of a quadrilateral may be less than 360°, as far as we know. In fact, we don't know whether or not there *are any* quadrirectangles.

THEOREM XII. *If there exists* **one** *triangle whose angle-sum is 180° then there exists a quadrirectangle.*

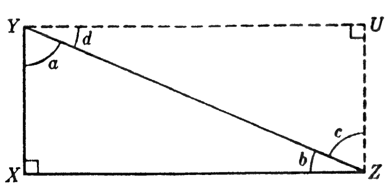


FIG. 230

Proof. By Theorem XI, there exists a right triangle XYZ whose angle-sum is 180°. Construct a congruent right triangle (as in Fig. 230) on the other side of YZ with right angle at U. Then $a = c$, $b = d$. But

(1) $$a + b = 90°$$

and

(2) $$c + d = 90°$$

since $X + a + b = 180°$ and $U + c + d = 180°$. Substituting d for b in (1) we get $a + d = 90°$ or $Y = 90°$. Substituting b for d in (2) we get $c + b = 90°$ or $Z = 90°$. Therefore $XYUZ$ is a quadrirectangle. This completes the proof.

THEOREM XIII. *If $ABCD$ is a quadrirectangle and (Fig. 231) if at any point E on AD we construct EF perpendicular to AD, then $ABFE$ is a quadrirectangle.*

Proof. Now, we have $p + q = 180°$, and $r + s = 180°$, although we do not know at present whether or not r and s are right angles. By hypothesis, $A + B + C + D = 360°$. Hence $A + B + C + D + p + q + r + s = 720°$. Or, $(A + B + r + p) + (s + q + C + D) = 720°$. Now, if one of these parentheses were less than half of 720° the other would have to be more than half of 720°. But each parenthesis is the angle-sum of a quadrilateral and, by Theorem X, cannot be more than 360°. Hence each of them is exactly 360°. Therefore $r = 90°$, which completes the proof.

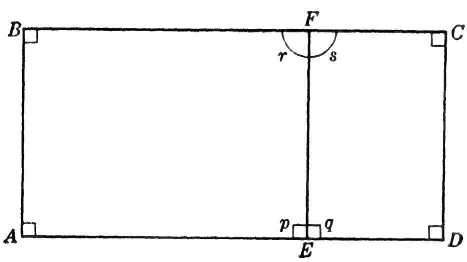


FIG. 231

Intuitively, this theorem provides a way of snipping off a smaller quadrirectangle from a larger one.

THEOREM XIV. *If $ABCD$ is a quadrirectangle and if a quadrirectangle $CDEF$ is constructed on the other side of CD, the resulting figure $ABFE$ is a quadrirectangle (Fig. 232).*

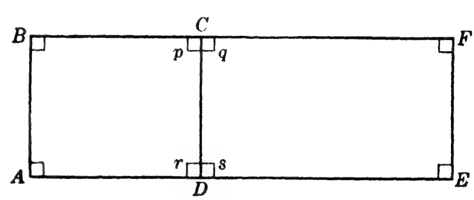


FIG. 232

Proof. Since p and q are right angles, BCF is a straight line. Similarly ADE is a straight line. Therefore $ABFE$ *is* a quadrilateral. The rest is obvious.

Intuitively, this means that when we put together two quadrirectangles on a common side, we get a larger quadrirectangle.

THEOREM XV. *If there exists* **one** *triangle whose angle-sum is 180°, then we can construct a quadrirectangle whose base and altitude * are greater than any given lengths x and y.*

Proof. By Theorem XII, there is *one* quadrirectangle $ABCD$. If this quadrirectangle happens to be large enough, we have nothing further to do; but its base and altitude may be much smaller than the given line-segments. In that case we can duplicate it by Theorem XIV as many times as necessary (as in Fig. 233) to make the base of the resulting quadrirectangle $ABMN$ larger than the given length x. Then by Theorem XIV, we may duplicate $ABMN$ as many times as necessary (Fig. 233) to make the altitude of the resulting quadrirectangle $BMPQ$ greater than the given length y.

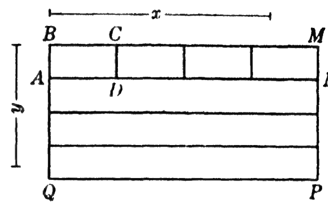


FIG. 233

THEOREM XVI. *If there exists* **one** *triangle whose angle-sum is 180° then we can construct a quadrirectangle whose base and altitude have exactly any given lengths b and a.*

Proof. By Theorem XV, we can construct a quadrirectangle $ABCD$ whose base and altitude are greater than b and a respectively. Lay off $AE = b$ on AD. Construct EF perpendicular to AD. By Theorem XIII, $AEFB$ is a quadrirectangle. Lay off $AG = a$ on AB. Construct GH perpendicular to AB. By theorem XIII, $AGHE$ is a quadrirectangle, and its base and altitude are as required.

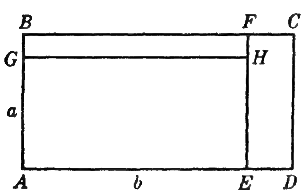


FIG. 234

Intuitively, we make a bigger quadrirectangle and then snip off one of the right size.

THEOREM XVII. *If there exists* **one** *triangle whose angle-sum is 180° then every* **right** *triangle has an angle-sum of 180°.*

Proof. Let the lengths of the legs of *any* given right triangle be a and b. By Theorem XVI, we can construct a quadrirectangle

* By a base and an altitude of a quadrirectangle, we mean any two adjacent sides One must not assume, however, that opposite sides of a quadrirectangle are equal

ADBC with base and altitude equal to b and a respectively (Fig. 235). Draw *BA*. Then triangle *ABC* is congruent to the given right triangle by Theorem I. Now we have $(C + p + r) + (q + s + D) = 360°$ and if one of these parentheses were less than half of 360° the other would have to be more than half of 360°. But each parenthesis is the angle-sum of a triangle which, by Theorem VII, cannot be more than 180°. Hence *ABC* is a right triangle whose angle-sum is 180°. But the given right triangle is congruent to triangle *ABC*. Hence the angle-sum of the given right triangle is 180°. This completes the proof.

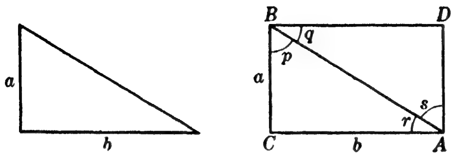


FIG. 235

THEOREM XVIII. *If there exists* **one** *triangle whose angle-sum is 180°, then* **every** *triangle has an angle-sum of 180°.*

Proof. Let *ABC* be any given triangle. From one of the vertices, say *B*, we can draw the perpendicular *BD* to the opposite side (Fig. 236). Then $A + p + r = 180°$ and $C + q + s = 180°$ by Theorem XVII. Hence $A + C + (p + q) + (r + s) = 360°$. But $r + s = 180°$. Hence, $A + C + (p + q) = 180°$ or $A + C + B = 180°$.

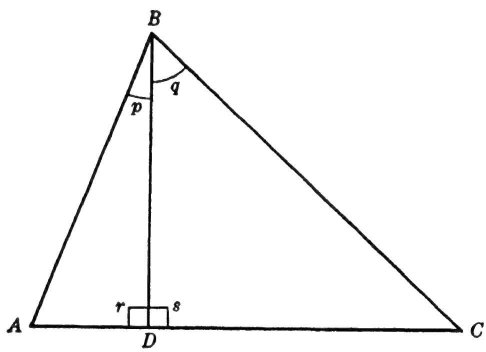


FIG. 236

THEOREM XIX. *If there exists* **one** *triangle whose angle-sum is less than 180°, then* **every** *triangle has an angle-sum less than 180°.*

Proof. By Theorem VII, the angle-sum of every triangle is $\leqq 180°$. By Theorem XVIII, if the angle-sum of *one* triangle were 180° then the angle-sum of *every* triangle would be 180°, contrary to the hypothesis. Thus the angle-sum of every triangle must be $< 180°$.

By virtue of Theorem XIX, in order to prove that the angle-sum of every triangle is less than 180° in Lobachevskian geom-

etry, it will be sufficient to show that there is *one* triangle with angle-sum less than 180°.

Remark. All the previous theorems are common to Euclidean and Lobachevskian geometry since their proofs are based only on those postulates which are made in both geometries; that is, no parallel postulate of any kind has been involved. If we now introduced the Euclidean parallel postulate, we would get Euclidean geometry. We could then prove the following theorems (a), (b), and (c).

(a) *If two alternate interior angles made by a transversal to two lines are not equal, the lines are not parallel.*

Proof. One of the two unequal angles p and q (Fig. 237) is greater than the other; say $p > q$. Then at A construct angle r = angle q. Then AD is parallel to BC by Corollary IVa. But $p > r$, by substitution. Hence AD is different from AE and both lines pass through A. Hence AE cannot also be parallel to BC by the Euclidean parallel postulate.

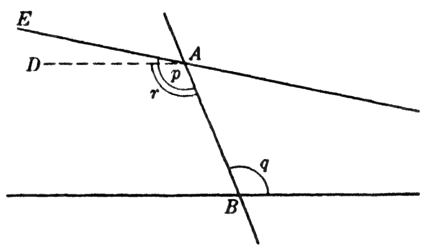


FIG. 237

(b) *If two lines are parallel, then any two alternate interior angles made by a transversal are equal.* (Converse of Corollary IVa.)

Proof. If there were a pair of unequal alternate interior angles, the lines could not be parallel, by (a).

(c) *The sum of the angles of any triangle is 180°.*

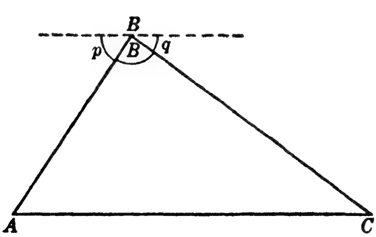


FIG. 238

Proof. Let ABC be any triangle. At B (Fig. 238), draw the line parallel to AC. Then by (b), $p = A$, $q = C$. Hence $A + B + C = p + B + q = 180°$, since $p + B + q$ is a straight angle.

This is the familiar theorem of Euclidean geometry.

Instead of making the Euclidean parallel postulate, *let us now*

make the Lobachevskian parallel postulate. We then get Lobachevskian geometry. We can now prove the following theorem.

THEOREM XX. *There exists a triangle whose angle-sum is less than 180°.*

Proof. Let l be a line and Q a point not on l. Draw QP perpendicular to l (Fig. 239) and l' perpendicular to QP at Q. Then l' is parallel to l by Corollary IVa. By the Lobachevskian parallel

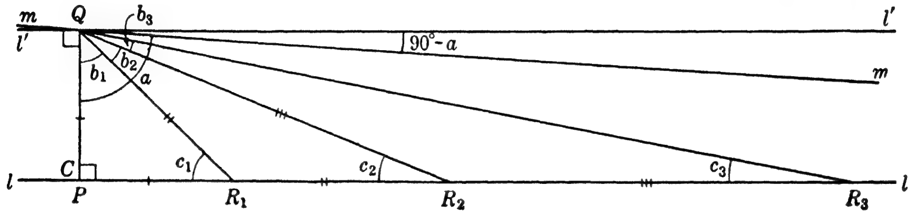


FIG. 239

postulate there exists another line m different from l', also parallel to l and passing through Q. By Theorem III, m cannot be perpendicular to QP since l' is perpendicular to QP. Hence one of the angles a made by m and QP must be less than 90°. Lay off $PR_1 = QP$ on l on the same side of the transversal QP as the acute angle a. Draw QR_1. Then $b_1 = c_1$ by Corollary Ia. By Theorem VIII, $c \geqq b_1 + c_1$. Substituting c_1 for b_1, we have $c \geqq c_1 + c_1$, or $c \geqq 2\,c_1$, or $c_1 \leqq c/2$. But $c = 90°$. Hence $c_1 \leqq 45°$. Now lay off $R_1R_2 = QR_1$ and draw QR_2. Then $b_2 = c_2$ by Corollary Ia, and $c_1 \geqq b_2 + c_2$ by Theorem VIII; thus $c_1 \geqq 2\,c_2$, or $c_2 \leqq c_1/2$, or $c_2 \leqq (1/2) \cdot (c/2)$, or $c_2 \leqq c/2^2$, or $c_2 \leqq 22.5°$. Repeating this argument we get $c_3 \leqq 11.25°$, and so on. After n such steps, we have $c_n \leqq c/2^n$, or $c_n \leqq 90°/2^n$. By the remark preceding Theorem VII, there exists a natural number n such that $90°/2^n$ is less than the positive quantity $90° - a$, no matter how small $90° - a$ may be. (For example, if a were 89°, $90° - a$ would be 1°, and we would have $c_4 \leqq 5.625°$, $c_5 \leqq 2.8125°$, $c_6 \leqq 1.40625°$, $c_7 \leqq .703125°$; thus $c_7 < 1°$ and $n = 7$ steps would be satisfactory in this example.) Hence, in triangle PQR_n, $\measuredangle PR_nQ = c_n < 90° - a$. Now, for every natural number n, $\measuredangle PQR_n < a$; for m does not meet l while QR_n does, so that QR_n lies below m. Finally $\measuredangle P = 90°$. Hence the angle-sum of triangle $PQR_n = \measuredangle P + \measuredangle PR_nQ + \measuredangle PQR_n <$

$90° + (90° - a) + a$; that is, the angle-sum of triangle PQR_n is less than 180°. (In our illustrative example, $\measuredangle a = 89°$. Hence $\measuredangle PQR_7 < 89°$, $c_7 = \measuredangle PR_7Q < 1°$, $\measuredangle P = 90°$ and therefore the angle-sum of triangle PQR_7 is $< 180°$.) This completes the proof.

THEOREM XXI. *The angle-sum of every triangle is less than 180°.*

Proof. Follows immediately from Theorems XX and XIX.

Remark. The proofs in this section involve some theorems which were taken intuitively but which could have been proved logically from a complete set of postulates. In particular, they are riddled with inferences drawn from the diagrams or pictures. The derivation of Lobachevskian geometry, or, for that matter, Euclidean geometry, from postulates by strict logic without any pictorial inference is a difficult task which is best left for advanced courses. Your high school Euclidean geometry text took intuitively many things such as we slid over here. The references at the end of the chapter will provide places where one can find sound logical treatments of Euclidean geometry (especially, Veblen, and Forder).

EXERCISES

Prove, in Lobachevskian geometry, that:

1. There exists no quadrilateral with an angle-sum of 360°.
2. If the angles of one triangle ABC are respectively equal to the angles of

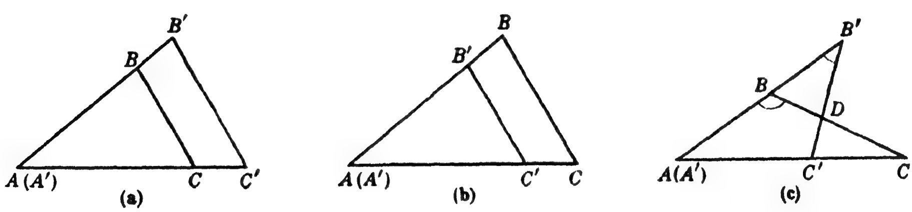


FIG. 240

another triangle $A'B'C'$ (that is, if $A = A'$, $B = B'$, and $C = C'$), then the two triangles are congruent.* (Hint: Suppose the contrary. Lay off one triangle on

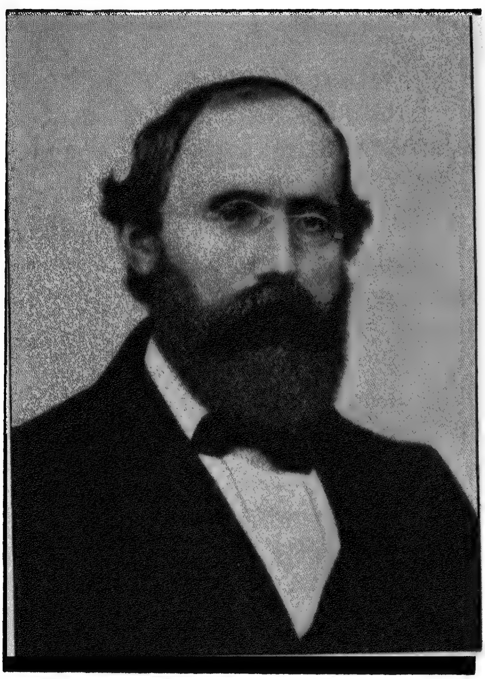
* It follows that in Lobachevskian geometry there are no similar but non-congruent figures. Hence if space were really Lobachevskian it would be fair to say that any picture of your sweetheart fails to do her or him justice, unless the picture were life-size.

the other so that $\measuredangle A$ and $\measuredangle A'$ coincide. Show that Fig. 240c cannot occur and show that the resulting quadrilateral $BCC'B'$ must have an angle-sum of 360° contrary to exercise 1.)

3. If in a quadrilateral $ABCD$ we have $AD = BC$ and $A = B = 90°$, then the angles D and C are equal but acute. (Hint: draw DB and AC.)

4. Is the theorem "if two distinct lines are both parallel to the same line, then they are parallel to each other" true in Lobachevskian geometry?

152. Riemannian geometry. Riemann (1826–1866) developed many non-Euclidean geometries, different from Lobachevski's. In the two classic ones, called single-elliptic and double-elliptic geometry respectively, one assumes instead of the Euclidean parallel postulate the following:



Georg Friedrich Bernhard Riemann
1826–1866, German

Riemannian parallel postulate. *There are no parallel lines.*

That is, any pair of lines must meet somewhere. Single-elliptic geometry is difficult to picture concretely, so we shall discuss here only the Riemannian geometry known as "double-elliptic geometry" which can be pictured easily. Double-elliptic geometry differs from Euclidean geometry not only by assuming the Riemannian parallel postulate instead of the Euclidean parallel postulate but also contradicts another Euclidean postulate, namely the one designated as Postulate 2 (section 149). Postulate 1, however, is retained, so that, in this geometry, two distinct points may have one or more lines passing through them. We shall refer to this double-elliptic geometry as **Riemannian geometry** hereafter, since it is the only one of Riemann's geometries that we shall discuss.

To show that this Riemannian plane geometry is just as consistent as Euclidean geometry we can make a model of a Riemannian "plane" in Euclidean space, as follows. Consider the surface of a sphere in Euclidean space; call it a **Riemannian plane.** Let us make a concrete interpretation of the abstract logical system called Riemannian plane geometry as follows. The undefined term "point" in Riemannian plane geometry shall be interpreted as meaning an ordinary (Euclidean) point on the surface of the sphere. The undefined term "line" in Riemannian plane geometry shall be interpreted as meaning a great circle on the sphere (that is, a circle formed by the intersection of the sphere with a Euclidean plane passing through the center of the sphere). Then all Riemann's postulates can be seen to be verified (Fig. 241). If 2 points happen to be diametrically opposite on the sphere then there are many "lines" (great circles) through them. For example, all the meridians of longitude lie along great circles through the North and South poles. Clearly all "lines" (great circles) intersect.

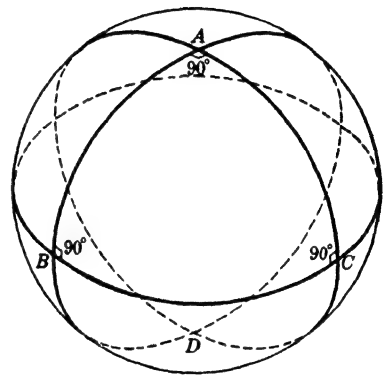


FIG. 241

It follows that Riemann's geometry is as consistent as Euclid's. For if Riemann's geometry involved any contradiction, the fact that we can have a model of Riemann's geometry in a Euclidean space would imply that the same contradiction must exist in the geometry of great circles on a sphere in Euclidean geometry.

As for applying to reality, the answer is much as before. In Riemann's system the sum of the angles of any triangle is more than 180°. For example, on the earth's surface, a triangle formed by the equator, the Greenwich meridian (0° longitude) and the meridian of 90° West Longitude, has in it three right angles (Fig. 241). But measurement does not reveal whether the sum of the angles of any physical triangle is exactly 180° or 180.000002° (or 179.99999876°). Therefore this test does not enable us to say that Riemannian geometry does not fit reality. In fact, suppose that what we think of as a plane were really a spherical surface of

tremendously large radius. Then any "small" region of the surface would appear to be flat, just as small regions on the earth's surface appear to be flat within the limits of our ability to measure although the earth is really approximately spherical. Thus surveyors, measuring small tracts of real estate, use the formulas of Euclidean plane geometry although they know that the earth is not flat, because in the small region of the earth's surface in which they are operating the differences between the two systems would not be measureable. Now "small" is a relative term and we can easily imagine a sphere of radius so large that a region on its surface whose size was greater than the range of our actual astronomical experience would appear to be flat within the limits of our ability to measure. Thus it might be that the geometry of "actual" space is really Riemannian. In fact, the discrepancies between the theorems of Euclidean and Riemannian geometry, in any limited region such as that of our actual experience, can be made too small to measure by merely supposing that the radius of the sphere is large enough.

On the other hand we must not suppose that the surface of a Euclidean sphere is the only possible concrete interpretation of Riemannian geometry; just as we must not suppose that the interior of a Euclidean circle is the only possible concrete interpretation of Lobachevskian geometry. In fact the whole point is that, as far as we can tell, the actual universe may be a concrete interpretation of Riemannian geometry. The spherical surface is introduced here merely to establish that Riemannian geometry is just as logically consistent as Euclidean.

Remark. On a flat surface (Euclidean plane), the shortest path joining 2 points is a straight line. It can be proved that on a spherical surface, the shortest path joining two points lies along the great circle passing through them. Hence, navigators on the ocean and aviators on long trips travel along great circles. Therefore it is not unnatural to expect great circles on a sphere to have properties analogous in many respects to the properties of straight lines in a (Euclidean) plane. Thus, in a sense Riemannian geometry is suggested by our practical experience. Of course, the geometry of the sphere was studied by the ancient Greeks. But it would not have occurred to them to interpret the undefined term "line" as meaning a great circle on a sphere, since

they thought of lines as having the usual intuitive meaning and not as undefined terms at all.

153. Conclusion. We have seen that other abstract mathematical sciences than Euclidean geometry may apply equally well to the physical world. We have looked at two forms of non-Euclidean geometry, Lobachevski's and Riemann's. There are other geometries besides these which are also different from Euclid's. One such geometry has actually been found more convenient than Euclidean geometry in Einstein's physical theory of relativity. However, the discrepancies between the results of Einstein's theory and those of Newton's physical theory which is based on Euclidean geometry are so small in size that they cannot be measured except in connection with astronomical distances. Thus for practical work, on the earth, we continue to use Euclidean geometry and Newtonian physics since they are much simpler to handle.

However, this does not imply that non-Euclidean geometry is unimportant. In fact, it was the attempt to prove the Euclidean parallel postulate which led to the invention of non-Euclidean geometries which, in turn, led to a deep reconsideration of the nature of mathematics and forced mathematicians to arrive at a clear understanding of the difference between abstract mathematical sciences or Pure Mathematics and concrete interpretations or Applied Mathematics. It led to the overthrow of the idea that our assumptions are self-evident truths. It led to the understanding of the difference between truth and validity in connection with scientific theories. For, if one of these alternative geometries (Euclid's, Lobachevski's, or Riemann's) is actually *true* of the physical world, then the others are false, since they contradict each other. But they do not contradict themselves, and each of these logically consistent geometries seems to fit the real world as well as the others. Hence, one can no longer say that the axioms of Euclid are "absolute truths." It led to the wholesome practice (among scientists) of examining closely and challenging the underlying assumptions upon which our theories are based.

In stirring up these questions, and hence leading to a clear understanding of the nature of pure mathematics and applied

mathematics, the study of non-Euclidean geometries performed its greatest service. These matters were discussed in Chapter II, sections 8 and 9. Now that the reader has more mathematical background, we shall take them up again, and in greater detail, in the next two chapters.

REFERENCES

E. T. Bell, *The Queen of the Sciences*, G. E. Stechert, N. Y., 1938.

E. T. Bell, *The Search for Truth*, Reynal and Hitchcock, N. Y., 1934.

H. S. Carslaw, *Elements of Non-Euclidean Plane Geometry*, Longmans Green, London, 1916.

J. L. Coolidge, *Non-Euclidean Geometry*, Oxford Univ. Press, London, 1927.

H. G. Forder, *Euclidean Geometry*, Cambridge Univ. Press, Cambridge, 1927.

D. Hilbert, *Foundations of Geometry* (translated by Townsend) Open Court, Chicago, 1921.

H. P. Manning, *Non-Euclidean Geometry*, Ginn, Boston, 1901.

H. Poincaré, *The Foundations of Science*, The Science Press, N. Y., 1913.

J. W. Young, *Fundamental Concepts of Algebra and Geometry*, Macmillan, N. Y., 1911.

J. W. A. Young, *Monographs on Modern Mathematics*, Longmans Green, N. Y., 1911. (Chapters by O. Veblen and F. S. Woods.)

Chapter XVII

TWO SIMPLE MATHEMATICAL SCIENCES

154. Introduction. Throughout the book we have called attention repeatedly to the nature of abstract mathematical sciences and their concrete interpretations or applications, and we have discussed many illustrations of these things. Beginning with Chapter II, sections 8 and 9, we have seen that if one attempts to treat any subject matter logically one is inevitably led to base the treatment on some unproved assertions, called postulates, and some undefined terms; otherwise one becomes enmeshed in circular reasoning or circular definitions or both. When we do begin with unproved postulates and undefined terms, and we proceed to define new terms and prove new assertions by deductive logic we obtain an abstract mathematical science or abstract logical system, and we have been engaged in postulational thinking. Thus an abstract mathematical science does not refer to any particular concrete subject matter. A concrete interpretation or application arises when we assign some meanings to our undefined terms. Of course, our postulates and undefined terms are usually suggested by experience and are chosen with some application in mind. But we have seen that one of the advantages of abstract mathematical sciences is that they may well be applicable to several different subject matters or concrete interpretations. See, for instance, the example of section 19 and the trivial examples 1 and 2 of section 8. Some of our work in algebra and geometry has been done in the form of abstract mathematical sciences. But we were unable to do all our work here on a strict logical basis because algebra and geometry are complicated abstract mathematical sciences based on many postulates. In this chapter we shall study two simple examples of abstract mathematical sciences and their concrete interpretations each of which is based on only a few postulates. We shall

examine more closely some of the characteristics of postulational thinking.

155. Groups. We shall discuss here a simple but very important example of an abstract mathematical science and some of its concrete interpretations. Let us take as undefined terms a set G of undefined objects, called **elements** of G, which we denote by small letters $a, b, c, \cdots$, and an undefined operation denoted by $\circ$, whose properties will be given in the postulates. The significance of these postulates will become clearer as soon as some concrete interpretations are discussed, immediately below.

We now assume the following postulates.

I. To every pair of elements a and b of G, given in the stated order, there corresponds a definite (unique) element of G, denoted by $a \circ b$. **(Law of closure for the operation $\circ$.)**

II. If a, b, c, are any elements of G, then $a \circ (b \circ c) = (a \circ b) \circ c$. **(Associative law for the operation $\circ$.)**

III. There exists a unique element of G, denoted by e, having the property that, if a is any element of G whatever, then $a \circ e = e \circ a = a$.

DEFINITION. *The element e is called the **identity** element.*

IV. To each element a of G there corresponds a unique element of G, denoted by a', having the property that $a \circ a' = a' \circ a = e$.

DEFINITION. *The element a' is called the **inverse** of a.*

A set G of elements with an operation satisfying these postulates is called a **group.** Note that the word "group" is used here in a technical sense, not in the everyday sense as a synonym of set, collection, or assemblage. The group G is called **commutative** or **Abelian** (after Abel) if it also satisfies the following postulate.

V. If a and b are any elements of G, then $a \circ b = b \circ a$. **(Commutative law for the operation $\circ$.)**

Postulate V is not a logical consequence of the other postulates as may be seen below from the fact that some concrete interpretations of Postulates I, II, III, IV, namely the sixth and sev-

enth interpretations below, do not satisfy V. For if V were a logical consequence of I–IV then V would have to be true whenever I–IV were true. We shall now exhibit several different concrete interpretations of the postulates for a group.

First interpretation. Let the set G be the set of all integers, $0, \pm 1, \pm 2, \cdots$, and let the operation $\circ$ be addition. Then Postulate I is satisfied since, to every pair of integers a and b, given in the stated order, there corresponds a definite (unique) integer, denoted by $a + b$. Postulate II is satisfied since if a, b, c, are any integers, $a + (b + c) = (a + b) + c$. Postulate III is satisfied since the integer 0, and no other integer, has the property that $a + 0 = 0 + a = a$ for any integer a. Hence, the identity element e in this interpretation is zero. Postulate IV is satisfied since to each integer a there corresponds a unique integer, namely $(-a)$, such that $a + (-a) = (-a) + a = 0$. Hence the inverse of a in this interpretation is the negative of a. Thus we have a group. In fact, postulate V is also satisfied since $a + b = b + a$ for any integers a and b. Hence, the set of integers with the operation $+$ form a commutative group in which 0 is the identity element, and $-a$ is the inverse of a.

Second interpretation. Let G be the set of all positive rational numbers (fractions). Let the operation $\circ$ be multiplication. Then if 1 is taken as the identity element and $1/a$ as the inverse of a, all the postulates for a commutative group are satisfied. Postulate I is satisfied since if a and b are any positive rational numbers there is a unique positive rational number $a \cdot b$. Postulate II is satisfied since if a, b, c are any positive rational numbers then $a \cdot (b \cdot c) = (a \cdot b) \cdot c$. Postulate III is satisfied since $1 \cdot a = a \cdot 1 = a$ for any positive rational number a. Postulate IV is satisfied since for each positive rational number a, there is a positive rational number $1/a$ such that $a \cdot (1/a) = (1/a) \cdot a = 1$. Postulate V is satisfied since $a \cdot b = b \cdot a$ for any positive rational numbers a and b.

However, if we interpreted G to be the set of all positive *integers* and the operation to be multiplication, all the postulates except IV would be satisfied. But IV would not, since the positive integer 3, for example, would have no inverse. That is, there is no positive integer x such that $3\,x = x\,3 = 1$.

Exercise 1. Would all five postulates for a commutative group be satisfied if G is taken to mean the set of positive integers and the operation is addition? Explain.

Exercise 2. Show that if G is the set of all even integers, the operation $\circ$ is addition, 0 is the identity element, and $-a$ is the inverse of a, then we have a commutative group. Write out all 5 postulates replacing the undefined terms by their meanings in this interpretation.

Third interpretation. Let G be the set of all rational numbers and let $\circ$ be addition. Let 0 be taken as the identity element and $-a$ as the inverse of a.

Exercise 3. Verify that the postulates for a commutative group are satisfied in this interpretation. Write out all the postulates replacing the undefined terms by their meanings in this concrete interpretation.

Exercise 4. Would all five postulates be satisfied if we let G be the set of all positive rational numbers and the operation is addition? Explain.

Fourth interpretation. Let G be the set of all rational numbers *except zero,* and let $\circ$ be multiplication. Let 1 be the identity element and let $1/a$ be the inverse of a.

Exercise 5. Verify that the postulates for a commutative group are verified in this interpretation. Write out all the postulates replacing the undefined terms by their meanings in this concrete interpretation.

Exercise 6. Would all five postulates be satisfied if we let G be the set of all rational numbers *including zero,* and the operation is multiplication? Explain.

Fifth interpretation. Consider the dial in Fig. 242. Let the elements of G be the numbers 0, 1, 2, 3, 4, alone. Let the operation $\circ$ be "addition" understood in the following sense: the "sum" $2 + 1$ shall mean the number to which the dial hand points after rotating in the indicated direction (see Fig. 242) through two spaces starting from 0 and then through 1 space. Thus $2 + 1 = 3$. Zero is interpreted to mean do not rotate at all.

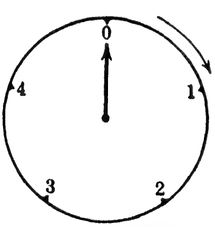


FIG. 242

In this way we get the "addition table":

$0+0=0$	$1+0=1$	$2+0=2$	$3+0=3$	$4+0=4$
$0+1=1$	$1+1=2$	$2+1=3$	$3+1=4$	$4+1=0$
$0+2=2$	$1+2=3$	$2+2=4$	$3+2=0$	$4+2=1$
$0+3=3$	$1+3=4$	$2+3=0$	$3+3=1$	$4+3=2$
$0+4=4$	$1+4=0$	$2+4=1$	$3+4=2$	$4+4=3.$

Let 0 be the identity element. From the addition table we see that the inverse of 2 is 3 (since $2 + 3 = 3 + 2 = 0$), the inverse of 4 is 1, (since $4 + 1 = 1 + 4 = 0$), and so on. Thus all the postulates except II can be verified easily by examining the table. The verification of Postulate II is tedious since one has to examine all possible triplets formed by numbers 0, 1, 2, 3, 4. For example to verify that $(2 + 4) + 3 = 2 + (4 + 3)$ we observe from the table that the left side is $1 + 3$ or 4 while the right side is $2 + 2$ or 4.

EXERCISES

7. Verify that $4 + (3 + 4) = (4 + 3) + 4$.

8. Verify that $2 + (3 + 4) = (2 + 3) + 4$.

9. Define $4 \cdot 3$ as $3 + 3 + 3 + 3$, $2 \cdot 4 = 4 + 4$, and so on; thus $4 \cdot 3 = 2$, $2 \cdot 4 = 3$, and so on. Make a "multiplication" table for the numbers on the dial above. Show that if we let the "elements of G" be the numbers 1, 2, 3, 4 (but not 0) and we let the operation $\circ$ be "multiplication," and we let 1 be the identity, then all the postulates for a commutative group are satisfied. Verify Postulate II only for 3 or 4 instances.

10. Make up an "addition table" like that of the fifth interpretation for the numbers 0, 1, 2, 3, 4, 5, on a similar dial. Show that the postulates for a commutative group are satisfied with addition as the operation and 0 as the identity. Verify Postulate II only for 3 or 4 instances.

11. Make up a "multiplication table" for the numbers on the dial of exercise 10, similar to that of exercise 9. Show that even though 0 is excluded, the elements 1, 2, 3, 4, 5 with "multiplication" as the operation $\circ$ and 1 as the identity do not satisfy all the postulates for a group. Which postulate is not satisfied? Verify Postulate II only for 3 or 4 instances.

Sixth interpretation. Let G be the set of six letters e, p, q, r, s, t. Let the operation $\circ$ be defined by the following "multiplication table" (or "operation table"):

$e \circ e = e$	$e \circ p = p$	$e \circ q = q$	$e \circ r = r$	$e \circ s = s$	$e \circ t = t$
$p \circ e = p$	$p \circ p = q$	$p \circ q = e$	$p \circ r = s$	$p \circ s = t$	$p \circ t = r$
$q \circ e = q$	$q \circ p = e$	$q \circ q = p$	$q \circ r = t$	$q \circ s = r$	$q \circ t = s$
$r \circ e = r$	$r \circ p = t$	$r \circ q = s$	$r \circ r = e$	$r \circ s = q$	$r \circ t = p$
$s \circ e = s$	$s \circ p = r$	$s \circ q = t$	$s \circ r = p$	$s \circ s = e$	$s \circ t = q$
$t \circ e = t$	$t \circ p = s$	$t \circ q = r$	$t \circ r = q$	$t \circ s = p$	$t \circ t = e$.

By inspection of this table we see that Postulates I and III are verified immediately, with e as the identity element. Similarly IV is obvious from the table; for example, the inverse of p is q

since $p \circ q = q \circ p = e$, and the inverse of r is r since $r \circ r = e$. Postulate II has to be verified by tediously inspecting all possible triplets of elements. For example, to verify that $p \circ (r \circ t) = (p \circ r) \circ t$ we observe from the table that the left side is $p \circ p$ or q and the right side is $s \circ t$ or q. Note that Postulate V is not satisfied; for example $p \circ r = s$ while $r \circ p = t$. Hence this is a group but not a commutative group. The construction of this table so as to satisfy postulates I through IV but not V may seem to be a remarkably ingenious feat; for if you try to make up such a table by random experiment, it will seem very difficult. In fact, however, the table given here was actually derived from the seventh interpretation below which is historically important. It can be shown that there is no simpler non-commutative group.

EXERCISES

12. What is the inverse: (*a*) of q; (*b*) of s; (*c*) of e; (*d*) of t?

13. Verify that $q \circ (s \circ t) = (q \circ s) \circ t$.

14. Verify that $q \circ (t \circ s) = (q \circ t) \circ s$.

Seventh interpretation. Let the numbers 1, 2, 3 represent volumes 1, 2, and 3 of a certain 3-volume work, respectively, standing in line on a shelf. Let us remove the books from the shelf and replace them in the order 3, 1, 2. The effect has been to replace volume 1 by volume 3, 2 by 1, and 3 by 2. Let the symbol $\begin{pmatrix}123\\312\end{pmatrix}$ stand for the act of replacing volume 1 by volume 3, volume 2 by volume 1, and volume 3 by volume 2. Let us denote $\begin{pmatrix}123\\312\end{pmatrix}$ by the letter p. In the same way, the symbol $\begin{pmatrix}123\\231\end{pmatrix}$ stands for the act of replacing volume 1 by volume 2, volume 2 by volume 3, and volume 3 by volume 1. Denote $\begin{pmatrix}123\\231\end{pmatrix}$ by q. Similarly let

$$r = \begin{pmatrix}123\\213\end{pmatrix}, \quad s = \begin{pmatrix}123\\321\end{pmatrix}, \quad t = \begin{pmatrix}123\\132\end{pmatrix}, \quad e = \begin{pmatrix}123\\123\end{pmatrix}.$$

Thus e stands for the act of replacing the books in the same order as they are, that is, of leaving the volumes as they are. Let $p \circ r$ mean perform the act p, then the act r. Thus p replaces 1 by 3 and r replaces 3 by 3. Hence $p \circ r$ replaces 1 by 3. Similarly p

replaces 2 by 1 and r replaces 1 by 2. Hence $p \circ r$ replaces 2 by 2. Finally p replaces 3 by 2 and r replaces 2 by 1. Hence $p \circ r$ replaces 3 by 1. Thus $\begin{pmatrix}123\\312\end{pmatrix} \circ \begin{pmatrix}123\\213\end{pmatrix} = \begin{pmatrix}123\\321\end{pmatrix}$. Or, $p \circ r = s$.

Exercise 15. Verify that the "multiplication table" for this interpretation is exactly the same as for the sixth interpretation. Hence this is a group but not a commutative group.

Eighth interpretation. Let G be the set of all vectors beginning at a given point in the plane. Let the operation $\circ$ be the addition of vectors (section 122); that is, the sum of two vectors is the vector represented by the diagonal of the parallelogram (except

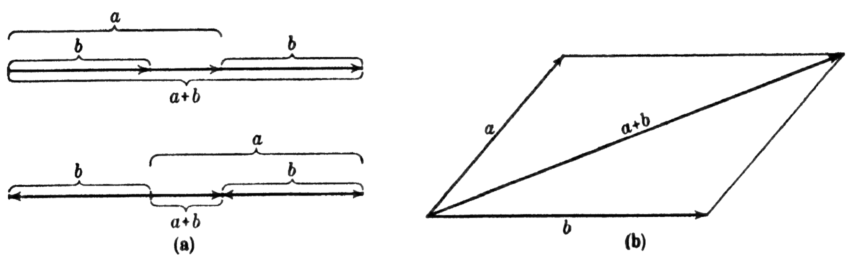


FIG. 243

when the two given vectors have the same or opposite direction in which case they are added by placing one after the other). Let the identity element be a vector of "length" zero. Let the inverse of any vector be a vector of the same length in the opposite direction.

Exercise 16. Verify that all the postulates for a commutative group are satisfied in this interpretation. To verify the associative law (Postulate II) is a rather complicated exercise in geometry.

The study of groups originated in situations like that of the seventh interpretation toward the end of the 18th and the beginning of the 19th centuries. These were studied by Lagrange (1736–1813), Abel, Galois, and others in connection with the solution of equations (see section 38). They also came to be of central importance in geometry. Groups are now a very fundamental study in higher mathematics and physics in many different connections. Although no practical application of groups was known 100 years ago, groups are now studied in connection with quantum physics.

It is clear that if we prove theorems in the abstract mathematical science based on Postulates I, II, III, IV, these theorems will apply automatically to all the different concrete interpretations of this abstract logical system. Hence, it is not necessary to prove these theorems over again in connection with each interpretation. This is one advantage of abstract or pure mathematics. It is a great unifying device which enables us to prove in one stroke results which apply to different subject matters. Furthermore, if a result is known in connection with one interpretation it may be discovered that the result depends only on our four postulates. This clarifies the nature of this result for our interpretation, and the result becomes automatically known for all other interpretations. In the next section we shall prove a few easy theorems about groups.

156. Some theorems about groups. We shall prove some theorems on the basis of Postulates I to IV only. Of course, we must be careful not to use anything not given in the postulates. We do not assume Postulate V. We must recall that a statement of the form $a = b$ means that a and b represent the same element of the group G. That is, a and b are merely different symbols or names for the same element of G. Hence, if $a = b$, either symbol may be substituted for the other freely. Thus, from $a = b$ we may conclude that $a \circ c = b \circ c$ or $c \circ a = c \circ b$, by substitution. Similarly if $a = b$ and $b = c$, we may conclude that $a = c$. Any inference of this sort will be referred to as "substitution."

THEOREM 1. *If a and b are any elements of G then $a \circ b'$ is an element of G.*

Proof. By Postulate IV, b' is an element of G. By Postulate I, $a \circ b'$ is an element of G.

THEOREM 2. *If a and b are any elements of G then $a' \circ b$ is an element of G.*

Proof. We leave this as an exercise.

THEOREM 3. *If $x = a' \circ b$ then $a \circ x = b$, all letters representing elements of G.*

Proof. If $x = a' \circ b$ then $a \circ x = a \circ (a' \circ b)$ by substitution. By Postulate II, $a \circ (a' \circ b) = (a \circ a') \circ b$. By Postulate

IV, $a \circ a' = e$. Hence $a \circ x = e \circ b$ by substitution. But $e \circ b = b$ by Postulate III. Hence $a \circ x = b$ by substitution.

THEOREM 4. *Conversely, if $a \circ x = b$ then $x = a' \circ b$, all letters representing elements of G.*

Proof. If $a \circ x = b$ then $a' \circ (a \circ x) = a' \circ b$ by substitution. By Postulate II, $a' \circ (a \circ x) = (a' \circ a) \circ x$. By Postulate IV, $a' \circ a = e$. Hence, $e \circ x = a' \circ b$ by substitution. By Postulate III, $e \circ x = x$. Hence $x = a' \circ b$ by substitution.

THEOREM 5. *If $a \circ b = a \circ c$ then $b = c$, all letters representing elements of G.*

Proof. If $a \circ b = a \circ c$ then $a' \circ (a \circ b) = a' \circ (a \circ c)$ by substitution. By Postulate II, we obtain $(a' \circ a) \circ b = (a' \circ a) \circ c$. By Postulate IV, $e \circ b = e \circ c$. By Postulate III, $b = c$.

THEOREM 6. *If a and b are any elements of G, then $(a \circ b)' = b' \circ a'$.*

Proof. By Postulate I and Postulate IV, $a \circ b$ and $b' \circ a'$ are elements of G. Hence by Postulate I, $(a \circ b) \circ (b' \circ a')$ is an element of G. By Postulate II, $(a \circ b) \circ (b' \circ a') = a \circ [b \circ (b' \circ a')] = a \circ [(b \circ b') \circ a']$. By Postulate IV and substitution, $(a \circ b) \circ (b' \circ a') = a \circ (e \circ a')$. By Postulate III, $a \circ (e \circ a') = a \circ a'$. By Postulate IV $a \circ a' = e$. Hence by substitution

$$(a \circ b) \circ (b' \circ a') = e. \tag{1}$$

By Theorem 4, if $h \circ x = k$ then $x = h' \circ k$. Applying this to (1) with $h = (a \circ b)$, $x = (b' \circ a')$ and $k = e$, we have $b' \circ a' = (a \circ b)' \circ e$. By Postulate III

$$b' \circ a' = (a \circ b)'$$

which is what we had to prove.

Let us see what one of these theorems means in various interpretations. For example, when interpreted in the sense of our first interpretation (section 155), Theorem 3 says "if $x = (-a) + b$ then $a + x = b$, all letters representing integers." When interpreted in the sense of the second interpretation,

Theorem 3 says "if $x = (1/a) \cdot b$ then $a \cdot x = b$, all letters representing positive rational numbers."

Many other theorems can be deduced from our four postulates. In fact, numerous books and articles in scientific journals are devoted to the study of the theory of groups.

EXERCISES

1. Restate Theorems 1 to 6 above in terms of each of the first two interpretations of section 155.

2. Prove that if $x = b \circ a'$ then $x \circ a = b$.

3. Prove that if $x \circ a = b$ then $x = b \circ a'$.

4. Prove that if $b \circ a = c \circ a$ then $b = c$.

5. Prove that $(a')' = a$.

6. Restate exercises 2–5 in terms of the first two interpretations of section 155.

157. A miniature geometry. In this section we shall study another abstract mathematical science and several concrete interpretations of it.

Let us take as undefined terms a set of undefined objects called **elements** and certain undefined classes of these elements called **l-classes.** Let us assume the following postulates. The significance of these postulates will become clear as soon as we examine some of the concrete interpretations of this abstract mathematical science.

I. There exists at least one l-class.

II. Not all elements belong to the same l-class.

III. Given two distinct elements, there is at most (not more than) one l-class containing them.

IV. Every l-class contains exactly three elements.

V. Given an l-class l and an element P not contained in l, there exists exactly one l-class l' containing P and having no element in common with l.

VI. Given an element P and an l-class l not containing P, there exists exactly one element P' contained in l such that no l-class contains both P and P'.

Note that we do *not* say "given any two distinct elements P

and P', there *is* an l-class containing both P and P'." Postulate II asserts only that there is no more than one such l-class; but there may be none. In fact, Postulate VI asserts that there are pairs of elements which are not contained in any common l-class.

Before proving any theorems in this abstract mathematical science, let us examine some of its concrete interpretations.

First interpretation. Let the "elements" be the nine letters A, B, C, D, E, F, G, H, I. Let the "l-classes" be the following nine classes of these elements:

$$\begin{array}{lll} (ABC), & (DBG), & (IBF), \\ (AEG), & (DEF), & (IEC), \\ (AHF), & (DHC), & (IHG). \end{array}$$

Exercise 1. Verify in detail that all 6 postulates are satisfied in this interpretation. Postulates I, II, IV are obviously satisfied. The others must be verified by systematic detailed inspection. For example, given the l-class (ABC) and the element D not contained in it, there is exactly one l-class, namely (DEF), containing D and having no element in common with (ABC). This verifies postulate V for this one instance. Verify Postulates III, V, and VI only for three cases each.

Second interpretation. The Mathematics department of a certain college has nine members whose names are found below, and is organized into nine committees. Let the "elements" be the members of the department and let the "l-classes" be the committees which are constituted as follows:

Committee on Curriculum:	Ames, Burk, Camp
Committee on Examinations:	Ames, Ellis, Green
Committee on Appointments:	Ames, Harris, Fisher
Committee on Student Affairs:	Davis, Burk, Green
Committee on Library:	Davis, Ellis, Fisher
Committee on Awards:	Davis, Harris, Camp
Committee on Graduate Studies:	Irving, Burk, Fisher
Committee on Public Relations:	Irving, Ellis, Camp
Committee on Finance:	Irving, Harris, Green.

Exercise 2. Verify that all six postulates are satisfied in this interpretation. (Hint: compare the first interpretation.)

Third interpretation. Let the "elements" be the nine points A, B, C, D, E, F, G, H, I in Fig. 244. Let the "l-classes" be the

nine triplets of these points which are joined by some straight line. Let us call these triplets of points "*lines*." Thus every line contains 3 points; that is, a "line" *is* a set of three points in this system. We may say that a point is "*on*" any line which contains it; a line is said to "*pass through*" any point on it. If two distinct points are on the same line l, they are said to be "*joined*" by l. If two distinct lines pass through (contain) the same point P they are said to "*meet*" or "*intersect*" at P. Two lines which do not meet (contain no common point) are called "*parallel*."

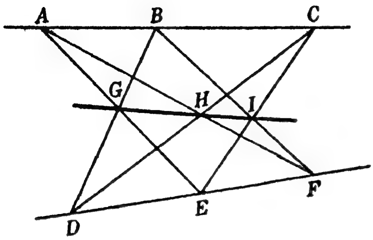


FIG. 244

Exercise 3. Verify that all 6 postulates are satisfied in this interpretation.

Since this interpretation is easiest to visualize we shall use the geometric terminology just introduced. The abstract mathematical science we are studying may be called a miniature geometry because of the geometric flavor of this interpretation. It may assist you in reading the next section to refer constantly to Fig. 244, remembering always that a "line" *is* a set of 3 points in this system, and that the streaks drawn in the figure are intended only as an aid to one's vision.

158. Some theorems in our miniature geometry. It will be convenient to restate our postulates in the geometric language introduced at the end of the previous section, as follows.

I. There exists at least one line.

II. Not all points are on the same line.

III. Given two distinct points, there is at most one line joining them.

IV. Every line passes through (*contains*) *exactly 3 points.*

V. Given any line l and a point P not on l, there exists exactly one line l' passing through P and parallel to l.

VI. Given a point P and a line l not passing through P, there exists exactly one point P' on l such that no line joins P and P'.

While Fig. 244 will be helpful in understanding the following

proofs, it must be remembered that we must not infer anything from our visual perception of the figure. That is, all arguments must be based strictly on the postulates, definitions, or previously proved theorems. Nothing else may be used except the principles of logic itself.

THEOREM 1. *There exists at least one point.*

Proof. Immediate consequence of Postulates I and IV.

THEOREM 2. *Not all lines pass through the same point.*

Proof. By Postulate I, there is a line l. By Postulate II, there is a point P not on l. By Postulate V, there is a line l' passing through P. But l does not pass through P. This proves the theorem.

THEOREM 3. *Two distinct lines have at most one point in common.*

Proof. If there were two distinct lines with *two* points in common, this would contradict Postulate III. Hence our theorem is proved.

THEOREM 4. *Exactly 3 lines pass through each point.*

Proof. (*a*) We prove first that *at least* 3 lines pass through each point. By Theorem 1, there exists at least one point; let A be any point. By Theorem 2, there exists a line not passing through A. By postulate IV, this line consists of 3 points, none of which can be A; call them D, E, F, say (Fig. 244). By Postulate VI, A is joined by lines to exactly 2 points of the line DEF. By Postulate V, there exists one line through A parallel to DEF; call it ABC. This line ABC together with the two lines obtained in the previous step give us three lines through A.

(*b*) We now prove that through A there pass *at most* three lines. Suppose there were a fourth line x through A. Either x is parallel to DEF or it meets DEF. Now x cannot be parallel to DEF, for if it were, we would have two lines, ABC and x, both through A and both parallel to DEF, contrary to Postulate V. On the other hand x cannot meet DEF. For A is already joined by lines to two points of DEF. Now x cannot join A to a third point of DEF since that would contradict Postulate VI; and x

cannot join A to either of the points of DEF to which A is already joined because that would contradict Postulate III. Hence, in any case, a fourth line x cannot exist. This proves (b). Our theorem follows from (a) and (b) immediately.

Many other theorems can be deduced from our six postulates, but their proofs are too difficult to be given here. One interesting theorem states that there must be exactly nine points, neither more nor less. This is a logical consequence of the postulates which could hardly be foreseen intuitively.

The miniature geometry which we have begun to study here was suggested by Fig. 244 which is called the Pappus configuration because of the following theorem of Euclidean geometry due to Pappus (about 340 A.D.): if A, B, C are points on one line and D, E, F are points on another line and if the pairs of lines AE and BD, AF and DC, BF and CE intersect in the points G, H, and I respectively then G, H, and I lie on the same straight line. The theorem of Pappus is really only a special case of a remarkable theorem of Pascal, a weak form of which may be stated as follows: if a hexagon $AECDBF$ is inscribed in a conic section so that all pairs of opposite * sides meet, then the three points of intersection of the pairs of opposite sides lie on the same straight line (Fig. 245). Note that a pair of lines is a special case of a conic

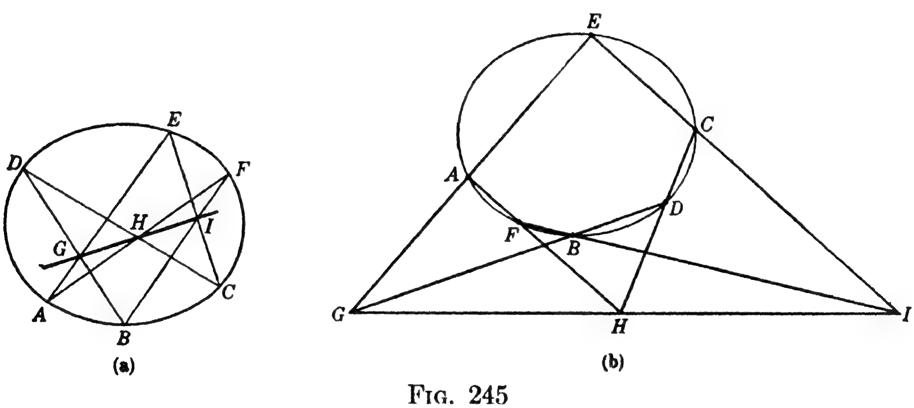


FIG. 245

section (see section 82). Pascal proved this theorem at the age of 16.

This miniature geometry exhibits a remarkable property

* A **hexagon** consists of six points joined in some order by six lines. The first and fourth, second and fifth, third and sixth lines are called **opposite** to each other.

known as duality. If in any statement we interchange the words "point" and "line," "on" and "through," "point of intersection" and "line joining," (corresponding changes in language being made to preserve grammar), the resulting statement is called the *dual* of the first, and the first is called the *dual* of the second. Now Postulates V and VI are duals of each other. Theorems 1, 2, 3, 4 are the duals of Postulates I, II, III, IV, respectively. Since the dual of every postulate is a correct proposition in this system, the dual of every theorem must therefore be correct as well. Thus the dual of any proposition already proved need not be given a separate proof. For example, the dual of the theorem which says that "there exist exactly 9 points" will state that "there exist exactly 9 lines."

159. Consistency and independence. To create an abstract mathematical science or abstract logical system, we have only to select some undefined terms, some unproved statements or postulates about them, and then define new terms in terms of the original undefined ones and prove new statements logically on the basis of the postulates. But it is essential that our postulates be consistent; that is, they must not be capable of leading to contradictory statements. How can we establish the consistency of a set of postulates? Clearly, the mere fact that we have deduced 100 or 10^{100} theorems without encountering any contradiction does not establish that we will *never* encounter one.

One way of proving the consistency of an abstract mathematical science is to exhibit a concrete interpretation of it; if concrete objects actually exist which satisfy the postulates then the postulates can not involve any contradiction. This faith in the logical consistency of the existing universe is so fundamental that it is hard to see how we could live without it. For example, the fact that the letters and l-classes of the first interpretation of section 157 exist and satisfy the postulates of our miniature geometry (section 157) establishes the consistency of that abstract mathematical science. Similarly the existence of the fifth or sixth interpretations of section 155 establishes the consistency of the postulates for a group.

However, such an exhibition of an interpretation using a finite number of concrete existing objects is not always possible. In

particular, it is impossible if one of the consequences deducible from the postulates is that the number of elements (such as points or numbers) in the science is infinite (as happens, for example in the case of the postulates for Euclidean geometry or in the case of Peano's postulates for the natural numbers).

Another method of proving the consistency of an abstract mathematical science is to give a concrete interpretation of it constructed within the framework of some other abstract mathematical science which is assumed to be consistent. In this way we established the consistency of the non-Euclidean geometries in Chapter XVI, assuming that Euclidean geometry is consistent. To establish the consistency of Euclidean geometry, we exhibit the concrete interpretation of its postulates afforded by the number-pairs, equations, etc., of analytic geometry (see section 87). This really establishes only that Euclidean geometry is as consistent as the number system. The consistency of the number system can be established if Peano's five postulates for the natural numbers are assumed to be consistent.*

To the question of the consistency of Peano's postulates no satisfactory answer is known. If none have seriously questioned their consistency, this seems to rest on a faith which goes beyond that in the logical consistency of the concretely existing universe. Intuitively, Peano's postulates arise from the counting process, which seems to be consistent; but this is hardly a consistency proof. (In fact, the counting process is always finite while Peano's postulates imply the existence of an infinite set.) The interpretation of Peano's postulates given by the finite cardinal numbers (section 148) establishes their consistency on the assumption of the consistency of certain logical operations involving sets; but that is no great progress, since these logical operations seem to be themselves more in need of justification than Peano's postulates.

Recently, consistency proofs for Peano's postulates for the natural numbers have been given, using a method, radically different from the method of giving a concrete interpretation, which depends on an analysis of the processes of deductive

* Properly speaking, besides the consistency of Peano's postulates, it is necessary to assume also the consistency of certain logical operations which are used in defining the real numbers in terms of the rational numbers. The validity of these logical operations has been seriously questioned in recent years by one school of mathematicians.

reasoning themselves by means of symbolic or mathematical logic. We cannot give here any more definite indication of the nature of these proofs. Suffice it to say that they also have their presuppositions, and the strength of these is such that no ultimate answer results to the question of the consistency of the elementary theory of the natural numbers. In any case, the subject of consistency has contributed to a rebirth of interest in symbolic or mathematical logic.

Another property of postulates for an abstract mathematical science which is aesthetically desirable, although not logically important, is independence. That is, if a postulate can be deduced from the remaining ones it is said to be dependent on them. In this case, there is no need to include the statement in question among the postulates. It is uneconomical and aesthetically unsatisfactory to assume something that need not be assumed because it can be proved. In the same spirit as this is an economical principle which is a good working rule in science; that is, we do not like to use many or complicated assumptions or hypotheses if few or simple ones will do just as well. This principle is known as **Occam's Razor.*** For example, we do not like to accept a mythological theory of natural phenomena which ascribes the cause of each phenomenon to the whim of a separate supernatural spirit, because we prefer a theory which can explain all these phenomena on the basis of a few general principles or "laws of nature" (postulates). To show that a postulate P is independent of the others, we have only to show that the others together with a postulate P', contradictory to P, form a consistent set. For if P could be deduced from the others, this latter set would be inconsistent since the resulting logical system would contain P and P'. Thus to show that Euclid's parallel postulate is independent of the other postulates of Euclidean geometry, we show that Lobachevskian geometry, which is based on all the postulates of Euclidean geometry (except the Euclidean parallel postulate) together with a parallel postulate contradictory to Euclid's, is consistent (see section 150). Similarly, to show that the commutative law is independent of the other four postulates for a commutative group, we exhibit concrete interpretations of

* After William of Occam (English, 14th century) who did not state it quite this way.

the other four postulates which do not obey the commutative law (see section 155, sixth interpretation).

Example 1. Let us show that Postulate IV of the miniature geometry of section 157 is independent of the others. Consider a new abstract mathematical science which differs from that of section 157 only in that Postulate IV is replaced by Postulate

IV′. Every l-class contains less than three elements.

To show that this system is consistent we merely have to exhibit a concrete interpretation of it, as follows.

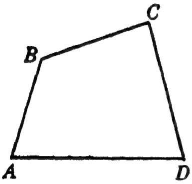


Fig. 246

Let the "elements" be the letters A, B, C, D and let the "l-classes" be the classes (AB), (AD), (BC), (DC). Then Postulates I, II, III, IV′, V, VI are all satisfied (Fig. 246). Another interpretation of this system is indicated in Fig. 246. Hence Postulate IV is independent of the others.

Exercise 1. Verify in detail that the concrete interpretation of example 1 satisfies Postulates I, II, III, IV′, V, VI.

Example 2. To show that Postulate V of section 157 is independent of the others, let the "elements" be interpreted as the letters A, B, C, D, E, F and the "l-classes" as the classes (ABC), (AEF), (BDF), (CDE). This interpretation satisfies the Postulates I, II, III, IV, VI, and (see Fig. 247)

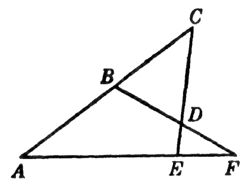


Fig. 247

V′. Every pair of l-classes have an element in common.

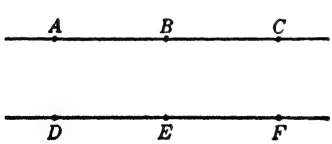


Fig. 248

Exercise 2. Verify the statements of example 2 in detail.

Exercise 3. Let the "elements" be the letters A, B, C, D, E, F, and the "l-classes" the classes (ABC), (DEF). This interpretation establishes the independence of one of the six postulates of section 157. Which one? Explain in detail. See Fig. 248.

In fact, each of the postulates of section 157 is independent of the others.

Remark. We can observe that all three interpretations of our miniature geometry of section 157 are essentially the same in

structure; they differ only in the names given to the elements and l-classes. If two interpretations are alike in this sense, they are called "isomorphic." That is, two interpretations are *isomorphic* if their terms can be placed in one-to-one correspondence in such a way that all correct statements * in one interpretation are translated into correct statements * in the other interpretation by merely substituting terms which correspond to each other in this correspondence (dictionary). As another example, the sixth and seventh interpretations of the abstract mathematical science of section 155 are isomorphic. If all possible interpretations of an abstract mathematical science are necessarily isomorphic, the science is called *categorical*. It can be proved that our miniature geometry is categorical. We shall not prove this here. The abstract science in section 155 is not categorical for we have some interpretations which have only a finite number of elements and others which have an infinite number of elements. This is sufficient to show that the system is not categorical, since the terms of a finite interpretation cannot be placed in one-to-one correspondence with the terms of an infinite interpretation at all, let alone in such a way that correct statements correspond to correct statements.

* Within the science.

Chapter XVIII

THE NATURE OF MATHEMATICS

160. Pure and applied mathematics. We return to the idea of an abstract mathematical science, or abstract logical system which has been illustrated many times throughout the book, perhaps most clearly in the preceding chapter. When we select some undefined terms, and a set of consistent postulates or assumptions about them, and proceed to define new terms in terms of the original undefined ones, and to deduce new statements logically from the preceding ones, we are creating an *abstract mathematical science.* If meanings are assigned to the undefined terms, then we have a *concrete interpretation* of our abstract mathematical science. The totality of all abstract mathematical sciences we call *Pure Mathematics.* The totality of concrete interpretations is called *Applied Mathematics.* Reasoning from postulates in this way is called *postulational thinking.*

This conception of mathematics is far broader than the common notion of what mathematics is, since it includes every subject in which you attempt to reason logically from explicitly recognized underlying assumptions, as you can hardly help doing, in all honesty, in any subject forming part of the search for truth.* For if you make an assertion in any subject whatever, you either assume it or must prove it. If you prove it, you deduce it from other assertions. If these are not to be assumed, they in turn must be deduced from still other statements. Sooner or later we must rest the entire structure on some unproved statements or postulates. We cannot prove all our statements unless we commit circular reasoning. Similarly if we consider any term used in our original statement, we either use it without definition or we define it. If we define it, we do so by expressing it in terms of other terms. These in turn are either left undefined or are

* The student is advised to review Chapter II, sections 8 and 9, at this point.

defined in terms of still other terms. Sooner or later we must rest all our definitions on some terms which are left undefined. We cannot define all our terms unless we use circular definitions. When we have organized our subject, whatever it may be, so that it begins with undefined terms and postulates and proceeds to define new terms and prove new statements, we have converted it into an abstract mathematical science. If you wish to attribute meanings (apprehended intuitively or some other way) to your undefined terms, you then obtain a concrete interpretation or application of your abstract science; that is, you then have a branch of applied mathematics.

Naturally, in passing from an abstract mathematical science to a concrete interpretation, we try to substitute such meanings for the undefined terms as will make true statements of the postulates. Needless to say, our abstract mathematical sciences are usually constructed with some application in mind. Of course whether or not a concrete interpretation really makes true statements of our postulates may not always be as certain as it was in sections 155 or 157, say. In the case of Euclidean geometry, we have seen that it is far from certain that our intuitive meanings of "dot" and "streak" for the undefined terms "point" and "line," respectively, really satisfy the Euclidean postulates. As we saw in Chapters II and XVI, the notion that our postulates are "self-evident truths" is not tenable. They are merely assumptions (whether or not they are suggested by experience, intuition, or creative imagination) and our theorems are deduced from the assumptions by logic.

But what of our rules of logic? Are they self-evident? There is hardly any doubt that most people even today would regard the rules of logic as self-evident and undeniable. But we have already burned our fingers in the fire of self-evidence in the case of the postulates of Euclidean geometry. A typical mathematical answer to the question of the self-evidence of logic itself is to formalize or formulate the rules of logic precisely and to say that we regard the logical rules as part of our assumptions. This leads to a careful study of formal, symbolic, or "mathematical" logic. Such studies go back to Leibniz, but the recent era in the development of symbolic logic begins with Boole, de Morgan, Peano, Frege, in the 19th century, and has been advanced by many con-

temporary workers such as Russell, Whitehead, Hilbert, Brouwer, Lukasiewicz, Tarski, and many others. Having taken the view that the rules of logic are assumptions and not self-evident truths, the possibility arises that *other* assumptions concerning logic may be consistent and useful. Thus arises the study of systems of logic different from the customary one; this is entirely analogous to the origin of non-Euclidean geometries. In fact, such logics have been studied in recent years, and have actually been found useful in connection with quantum physics. This work is at present going on.

As we saw in section 159, the study of mathematical or formal or symbolic logic has also been stimulated by the question of the consistency of the mathematical sciences. In addition, the study of the foundations of logical reasoning has received impetus from the occurrence of concealed inconsistencies in apparently harmless statements, like the following.

(*a*) "In a certain town, there is a barber who must shave all those people and only those people who do not shave themselves" (Bertrand Russell). This seems harmless, but does the barber shave himself? Clearly according to the rules, if he does, he mustn't; and if he doesn't he must. (This sounds like the famous words of Tweedledee in Lewis Carroll's *Through the Looking Glass:* "If it was so, it might be; and if it were so it would be; but as it isn't, it ain't. That's logic.") This barber cannot exist and not even Occam's Razor will shave him.

(*b*) "It is a rule that all rules have exceptions." If this rule has an exception, then there must be some rule without an exception; but if this rule has no exception, the statement is false.

(*c*) "This statement is false." If it's true, it's false; and if it's false, it's true.

Paradoxical statements like these indicate the necessity for a careful analysis of both logic and language. Perhaps the most paradoxical fact about modern mathematics is the fact that paradoxes have arisen within it. There is no general agreement concerning the resolution of some of these paradoxes. In fact, there are at present three different schools of thought concerning the logical foundations of mathematics and no reconciliation is in sight.

The King's advice to Alice to "begin at the beginning, and go

on till you come to the end; then stop" certainly does not apply to the development of mathematics. For the study of mathematics begins somewhere in its logical middle and progresses in two broad directions. One is the direction of further development of the logical consequences and applications of the various branches of the very ramified mathematical tree. The other is the direction of further delving into the underlying foundations or roots upon which the whole magnificent structure rests. Both directions of research are important and very much connected with each other. Every theorem of pure mathematics, from the elementary ones we have studied here to the most advanced ones available, is firmly cemented to the underlying assumptions by steps of strictly logical reasoning. There seems to be no end in either direction, and consequently no possibility that mathematics can ever become completely worked out, dead, embalmed in books, and devoid of interest. Since the 17th century new results in mathematics have been produced at ever increasing speed. In recent years, periodicals devoted exclusively to the publication of new mathematical research have contained thousands of articles annually on many different subjects. Mathematics is far from the dead subject that some students think it is. In fact, in modern times, it has had something of the appearance of the furious young man who mounted his horse and rode off in all directions. Needless to say, not all of these myriad contributions survive. But neither are they all completely lost. Every so often, someone synthesizes countless special results into a powerful general method. In fact, a subject sometimes begins its most interesting phase after it has been pronounced worked out and dead by a hasty coroner's jury. In any case, no mathematician is likely ever to voice Alexander's complaint concerning the scarcity of new lands to conquer.

161. Mathematics as a branch of human endeavor. The magnificent conception of mathematics as the study of all abstract logical systems or abstract mathematical sciences and their concrete interpretations or applications really justifies the statement that mathematics is basic to every subject forming part of the search for truth. In fact, mathematics, thus conceived, includes all subjects into which one injects logical structure. "To

mathematize a subject does not mean merely to introduce equations and formulas into it, but rather to mould and fuse it into a coherent whole, with its postulates and assumptions clearly recognized, its definitions faultlessly drawn, and its conclusions scrupulously exact." * That is, to mathematize any subject means simply to put it in the form of an abstract mathematical science. As remarked in Chapter II, and in the preceding section, this is unavoidable if one insists on organizing in a strictly logical way any subject matter at all. Realization of this should make us look into the assumptions underlying the things we assert. Examination of the underlying assumptions and the correctness of the reasoning by which we draw conclusions from them should serve to clarify our beliefs, to make us cognizant of the possibility of other tenable assumptions and other beliefs, and should therefore assist us on the road to tolerance, maturity, and wisdom. It is natural that, since antiquity, mathematics should have pointed the way toward these ideals of clarity since it has been comparatively free from the emotional confusion of prejudice and hatred; this is even more true since Lobachevski and others destroyed the sanctity of the notion that the axioms of Euclidean geometry constituted a self-evidently absolutely true description of physical space, a belief that was held with firm conviction by many people who scoffed at other "absolute truths." Mathematics today can still serve us as an ideal of intellectual honesty, of logical rigor and vigor and clarity, toward which to strive. Any applied science, for example, develops by evolutionary stages. First, the collection of data takes place; then the formation of hypotheses or postulates; then the deduction of the logical consequences of these postulates; then the checking of these consequences against the observed facts or data. If some consequence does not check with the facts, we attempt to revise the postulates and the process begins again. Ultimately the subject is put into the strict logical form of a mathematical science. Physics, chemistry, biology, and the social sciences have all followed this general evolutionary pattern, although biology and the social sciences are in the earlier stages of development, from this point of view.

As we have pointed out, mathematics is the backbone of all

* 15th Yearbook of the National Council of Teachers of Mathematics.

scientific (logical) subjects, whether physical, biological, social, or otherwise, and, as such, is and must continue to be of the utmost importance to the civilized world. We have already remarked that pure mathematics is usually developed with some application in mind. Thus most of elementary mathematics originated in the social needs of commerce, surveying, engineering, etc. However, it would be a tragic mistake to censor or prohibit research in pure science merely because there was no immediate practical application visible. History has shown us many times that the applications may come much later than the pure science. For example, the conic sections were studied in ancient Greece for their own sake, and came to be of overwhelming importance in physics only in the 17th century. Non-Euclidean geometry, developed for its own sake as pure mathematics early in the 19th century, was applied to the physical theory of relativity in the 20th century. Similarly, the theory of groups, developed for its own sake as pure mathematics in the 19th century, was applied to the physical theory of quantum mechanics in the 20th century. There can be no doubt that practical social needs have always had a great effect on the development of mathematics. Conversely, there can be no doubt that mathematics has had a great effect on the development of society, as, for example, in the case of the industrial revolution which would never have taken place without the technical applications of mathematics to engineering. Nor can there be any doubt that social conditions may well tend to stifle or stimulate scientific progress. Nevertheless, one should not lose sight of the fact that in pure mathematics we have a great structure of logically perfect deductions which constitutes an integral part of that great and enduring human heritage which is and should be largely independent of the perhaps temporary existence of any particular social or political conditions in any particular geographical location at any particular time. Euclid's geometry is far more important to us than the political ideas of Euclid's day. The enduring value of mathematics, like that of the other sciences and arts, far transcends the daily flux of a changing world. In fact, the apparent stability of mathematics may well be one of the reasons for its attractiveness and for the respect accorded it in a world wherein security is so elusive.

To a certain extent it is doubtless true, as has been maintained by Spengler in his *Decline of the West,* that the mathematics of any period is a good index of its culture. There is no doubt that much mathematical research has been definitely "in the air" and was not created out of whole cloth by any genius. This is indicated by the frequency with which similar results have been obtained by different men independently of each other. For example, no one would be greatly surprised by the invention of cheap television today although 50 years ago it would have been called impossible, and 500 years ago its inventor would have been burned at the stake. However, to conclude from this that the people who actually do the inventing are unimportant is an error that is more common than just. It is true that any two mathematicians will agree on the correctness of a mathematical theory, if they agree on the basic postulates. It may also be true, in at least some cases, that, if any given mathematician had not lived, his work might have been done by someone else, although perhaps much later; this opinion or conjecture has its defenders and its opponents. To this extent, mathematics may be independent of individual people. But to say that society or its needs "produced" a certain achievement is to ignore entirely that only one or a few individuals actually contributed to this achievement. To say that society or the human race "did it" is much like the remark of the baseball fan who says proudly that "we" won the ball game when he means that 9 athletes not even remotely connected with him won the ball game. In point of fact "society" does not always deserve such vicarious glory for it has sometimes seemed to do its best to censor, discourage, and hinder its best minds in every imaginable way.

We need not dwell here on the manifold materially practical applications of mathematics, since these are, in a general way, well known to the reader, except to point out that the contempt in which the business man or "man of action" sometimes holds the pure scientist is totally unjustified. Such a person overlooks completely the fact that his daily life is full of the results of pure science. For example, the amount of pure astronomy involved in the answer to the practical and important question "what time is it?" might well astound a "man of action." Of course the importance of pure science has been recognized by

big business in recent years and many large corporations support pure scientists in well-equipped laboratories in which they may pursue their researches, regardless of immediate applicability. Needless to say the practical by-products of this research have amply justified the investment from a purely commercial standpoint. We have therefore stressed the importance of mathematics in the world of ideas, since this aspect may not be quite so familiar. Mathematics has been aptly called both the queen and the handmaiden of the sciences. Mathematics, likewise, has had a tremendous influence on philosophy as well as on the sciences. For example, mathematics has successfully dealt with the problems of infinity and has contributed greatly to the solution of problems of logic, all of which are problems that originated in philosophy. More than that, however, mathematics has clarified the notions of validity and truth, of pure and applied mathematics, and has led, especially since the study of non-Euclidean geometries, to a clearer understanding of the nature of human knowledge.

While we have stressed the scientific characteristics of mathematics and its relations to the sciences, it must be said that mathematics has relations with the arts as well, and has also the characteristics of an art. By this statement we do not refer merely to the obvious relations of geometry to pictorial design or architecture, or to the relations between numbers and musical harmonies which were known to Pythagoras (see section 27). We mean rather that the creation of new mathematics is itself an art and that the contemplation of well wrought mathematical systems gives rise to genuine aesthetic satisfaction. It is easy to say that to construct an abstract mathematical science we have only to choose undefined terms and postulates and to define new terms and deduce new theorems. But what undefined terms, postulates, definitions shall we choose and which of the possibly infinite number of logical conclusions shall we deduce? Thus arises the problem of selection of materials to start with and the selection of a pattern or direction or mode of development of these materials. This is a problem of artistic composition requiring insight, taste, intuition, and creative imagination (that is, a problem of creative art), just as is the analogous problem of painting or musical composition, say, where one has to choose themes and

modes of developing and interweaving them out of an infinite variety of possibilities. In mathematics we have, of course, the partial criterion that the resulting logical structure be consistent and preferably applicable to something; just as in music we have the criterion that the resulting composition be pleasing to the ear and perhaps emotionally satisfying. But to meet these criteria in either mathematics or music may well require artistic genius. For example, the genius of Newton was needed for his beautiful logical theory of gravitation which seemed to introduce order and harmony into the apparently chaotic movements of the entire physical universe so well that his work remained unchallenged for over 200 years, and is still a valuable and elegant logical composition. As for the statement that the contemplation of mathematics gives aesthetic satisfaction, you must understand that it is assumed that the contemplator has the training or background needed for such appreciation. This is the case in any art. A beginner in a foreign language cannot appreciate a great piece of literature if his knowledge of vocabulary and grammar is fragmentary. An untrained person does not get all he might from hearing a beautifully wrought symphony and gets nothing at all from seeing it in musical print, if he reads the individual notes or symbols with difficulty, while a well-trained person can read a score and enjoy it. Similarly in mathematics, only when one passes the stage of reading the individual symbols with difficulty may one begin to see and appreciate in a mathematical composition the symphonic interplay of ideas, the economical simplicity of development, the subtle modulations and inversions of treatment, the inevitable sweep toward a climactic conclusion, and so on; just as in a musical composition one may see and appreciate the symphonic interplay of themes, the economical simplicity of development, the subtle modulations and inversions of treatment, the inevitable sweep toward a climax, and so on. The aesthetic satisfactions obtained in both instances by people whose background permits such appreciation are much akin. There can be no doubt that creative imagination has been needed for the invention of great mathematical works. In his *Philosophical Dictionary*, Voltaire wrote that ". . . there was far more imagination in the head of Archimedes than in that of Homer." Havelock Ellis, not a mathematician himself, says in

The Dance of Life that "it is here [that is, in mathematics] that the artist has the fullest scope for his imagination." The mathematician is not even confined to the "actual" world. He is free, in a sense, to explore all possible worlds.

Needless to say, mathematics is the nearest thing to an international language that the human race possesses. The symbolism of mathematics is the same the world over, and work done in one part of the world is likely to be taken up and advanced further thousands of miles away. From another viewpoint, mathematics may be considered *the* most distinctively human of all human activities. For, as far as we know, man's social instincts, etc., are shared with other animals, but reflective logical curiosity and the important trait of being able to accumulate, organize, and add to wisdom and pass it on to the next generation for further improvement is distinctly human. It is our logical minds that serve to bind the future with the past and to give us whatever hope we have for progress. For example, the suspension bridge on which you cross a river rests not only on its pillars but on the calculus of Isaac Newton, the analytic geometry of Descartes, the algebraic researches of many centuries, the geometry of the ancient Greeks and the primitive number-lore of our prehistoric ancestors. While our mathematical ideas change and progress, they do not do so by ignoring past achievements (although new developments often render past achievements obsolete). Real progress in any branch of the search for truth demands a degree of open-mindedness sufficient to combat both the unwise tendency to cling to old beliefs merely because they are old and the equally foolish tendency to embrace hastily every novelty merely because it is new. The history of mathematics and the physical sciences, for example, has shown many times the folly of both these policies. These cautions are obviously even more necessary in subjects which involve our emotions and in which the check of controlled experiments is either impractical or impossible. It is, however, worth pointing out that controlled experiments are seldom possible in astronomy. Yet in the course of many centuries, astronomical theories have been patiently and logically built up until astronomy may well be regarded as the most accurate and reliable branch of mathematical physics. To the appeal for rational thought it is sometimes objected that

when a house is on fire there is no time to conduct a dispassionate scientific investigation. This argument, however, does not justify the too common contempt for the reflective scientist and the corresponding adulation of the "man of action." It rather points to the desirability for calm scientific investigation *before* the need for hasty action arises. While man's logical ability has been most highly developed in connection with numbers and diagrams, may we not hope that, despite the obvious difficulties, man may yet become reasonable about more complicated and emotionally confused subjects? The standard of logical rigor set by mathematicians is a high ideal, but that scarcely justifies abandoning it. Man's faith in reason is so fundamental a need that to discard it lightly may be fairly regarded as a pathological symptom. But in order to strive toward this ideal, we must first understand and appreciate it. To give you such an understanding and appreciation, at least to some extent, has been one of the aims of this book. We hope that you are now able to see mathematics in its proper perspective as one of the greatest achievements of the human race.

REFERENCES

E. T. Bell, *The Development of Mathematics*, McGraw Hill, N. Y., 1940.

E. T. Bell, *The Handmaiden of the Sciences*, Williams & Wilkins, Baltimore, 1937.

E. T. Bell, *The Queen of the Sciences*, G. E. Stechert, N. Y., 1938.

E. T. Bell, *The Search for Truth*, Reynal & Hitchcock, N. Y., 1934.

A. A. Bennett & C. A. Baylis, *Formal Logic, A Modern Introduction*, Prentice Hall, N. Y., 1939.

S. Buchanan, *Poetry and Mathematics*, John Day, N. Y., 1929.

M. R. Cohen, *Reason and Nature*, Harcourt Brace, N. Y., 1931.

M. R. Cohen & E. Nagel, *Introduction to Logic and Scientific Method*, Harcourt Brace, N. Y., 1934.

A. Dresden, *Invitation to Mathematics*, Holt, N. Y., 1936.

Havelock Ellis, *The Dance of Life*, Houghton Mifflin, Boston, 1923. (Chapter on the Art of Thinking.)

C. J. Keyser, *Thinking About Thinking*, E. P. Dutton, N. Y., 1926.

C. J. Keyser, *Pastures of Wonder*, Columbia University Press, N. Y., 1929.

C. J. Keyser, *The Human Worth of Rigorous Thinking*, Columbia University Press, N. Y., 1925.

C. J. Keyser, *Mathematical Philosophy*, E. P. Dutton, N. Y., 1922.

C. J. Keyser, *Mathematics and the Question of the Cosmic Mind*, Scripta Mathematica, N. Y., 1935.

M. I. Logsdon, *A Mathematician Explains*, University of Chicago Press, Chicago, 1936.

H. Poincaré, *Foundations of Science*, The Science Press, N. Y., 1913.

B. Russell, *Mysticism and Logic*, Longmans Green, London, 1917.

B. Russell, *Our Knowledge of the External World*, Allen & Unwin, London, 1926.

B. Russell, *Introduction to Mathematical Philosophy*, Allen & Unwin, London, 1919.

B. Russell, *Principles of Mathematics*, Norton, N. Y., 1938.

D. E. Smith, *The Poetry of Mathematics and Other Essays*, Scripta Mathematica, N. Y., 1934.

A. Tarski, *Introduction to Logic*, Oxford University Press, 1941.

A. N. Whitehead, *Introduction to Mathematics*, Holt, N. Y., 1911.

J. W. Young, *Fundamental Concepts of Algebra and Geometry*, Macmillan, N. Y., 1911.

American Council on Education, "*Achievements of Civilization*":

#2. *The Story of Numbers.*

#4. *The Story of Our Calendar.*

#5. *Telling Time through the Centuries.* Washington, D. C., 1933.

Scripta Mathematica, a quarterly journal devoted to the philosophy, history, and expository treatment of mathematics. (Contains many articles suitable for the non-technical reader.)

APPENDIX

For the convenience of the student who decides to go on with mathematics after completing this course, we here append a brief treatment of some topics in trigonometry and algebra of a more technical nature, which are necessary preliminaries for a second year course in analytic geometry and calculus. We give only the barest minimum essentials. For further study the student is referred to text books on trigonometry and college algebra in which his reading will be best guided by the advice of his instructor.

162. Trigonometric identities. By definition (Fig. 249)

$\sin A = \frac{y}{r}$ while $\csc A = \frac{r}{y}$. Hence $\sin A \cdot \csc A = \frac{y}{r} \cdot \frac{r}{y} = 1.$

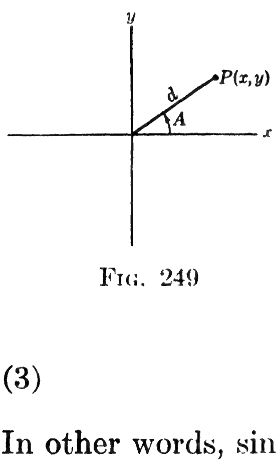


FIG. 249

Note that

(1) $$\sin A \cdot \csc A = 1$$

implies both $\sin A = \frac{1}{\csc A}$ and $\csc A = \frac{1}{\sin A}$. In the same way we deduce from the definitions (section 124) the statements

(2) $$\cos A \cdot \sec A = 1$$

and

(3) $$\tan A \cdot \cot A = 1.$$

In other words, $\sin A$ and $\csc A$ are *reciprocals* of each other as are $\cos A$ and $\sec A$, and $\tan A$ and $\cot A$.

Also, $\dfrac{\sin A}{\cos A} = \dfrac{\frac{y}{r}}{\frac{x}{r}} = \dfrac{y}{r} \cdot \dfrac{r}{x} = \dfrac{y}{x} = \tan A$. Hence,

(4) $$\tan A = \frac{\sin A}{\cos A}.$$

In the same way, we obtain

(5) $$\cot A = \frac{\cos A}{\sin A}.$$

From the fact that $x^2 + y^2 = r^2$ we obtain, by dividing both sides by r^2, the statement $\frac{x^2}{r^2} + \frac{y^2}{r^2} = 1$, or $\left(\frac{x}{r}\right)^2 + \left(\frac{y}{r}\right)^2 = 1$, or

(6) $$\cos^2 A + \sin^2 A = 1.$$

Dividing both sides of $x^2 + y^2 = r^2$ by x^2 we have $1 + \left(\frac{y}{x}\right)^2 = \left(\frac{r}{x}\right)^2$, or

(7) $$1 + \tan^2 A = \sec^2 A.$$

Dividing both sides of $x^2 + y^2 = r^2$ by y^2 we have $\left(\frac{x}{y}\right)^2 + 1 = \left(\frac{r}{y}\right)^2$, or

(8) $$\cot^2 A + 1 = \csc^2 A.$$

All the numbered equations are *identities*, since they are true for all values of the variable A for which they have meaning (section 32). They are called **fundamental identities.**

Remark. The expression $\sin^2 A$ means $(\sin A)^2$ not $\sin (A^2)$. We do not write $\sin A^2$ since this would cause confusion.

By using the fundamental identities we may verify a profusion of less obvious identities among the trigonometric functions, as in the following example. Note carefully what was said about the verification of identities in general in section 32.

Example 1. Verify the identity $(1 - \sin^2 A) \tan^2 A = \frac{1}{\csc^2 A}$.

The left member
$$\begin{aligned}(1 - \sin^2 A) \tan^2 A &= \cos^2 A \tan^2 A \quad \text{(by (6))}\\ &= \cos^2 A \frac{\sin^2 A}{\cos^2 A} \quad \text{(by (4))}\\ &= \sin^2 A,\end{aligned}$$

while the right member $\frac{1}{\csc^2 A} = \sin^2 A$ by (1). Since all the steps made are substitutions, they are all reversible and hence the given equation is an identity.

Example 2. Verify the identity $\tan A + \cot A = \csc A \sec A$.
The left member $\tan A + \cot A$

$$\begin{aligned}
&= \frac{\sin A}{\cos A} + \frac{\cos A}{\sin A} && \text{(by (4) and (5))}\\
&= \frac{\sin^2 A + \cos^2 A}{\sin A \cos A} && \text{(adding fractions)}\\
&= \frac{1}{\sin A \cos A} && \text{(by (6))}\\
&= \frac{1}{\sin A} \cdot \frac{1}{\cos A} && \\
&= \csc A \sec A && \text{(by (1) and (2)).}
\end{aligned}$$

Suggestion. Whenever an expression can be simplified by direct application of one of the fundamental identities (6), (7), or (8), do so at once. If no such step is obvious, express all functions in terms of sine and cosine alone by means of (1), (2), (4), (5), and simplify algebraically.

EXERCISES

1. Express each of the functions $\csc A$, $\sec A$, $\tan A$ and $\cot A$ in terms of $\sin A$ and $\cos A$ alone.

Verify each of the following identities:

2. $1 - \cos^2 A = \dfrac{1}{\csc^2 A}$.

3. $\dfrac{1}{1 - \sin^2 A} = \sec^2 A$.

4. $1 + \tan^2 A = \dfrac{1}{1 - \sin^2 A}$.

5. $\dfrac{1}{\cot^2 A + 1} = \sin^2 A$.

6. $\sin^2 A(\csc^2 A - 1) = \cos^2 A$.

7. $\cos^2 A(\sec^2 A - 1) = 1 - \cos^2 A$.

8. $(1 - \cos^2 A)\csc^2 A = 1$.

9. $(1 - \sin^2 A)\sec^2 A = 1$.

10. $(1 + \tan^2 A)(1 - \cos^2 A) = \sec^2 A - 1$.

11. $\csc^2 A \tan^2 A - 1 = \tan^2 A$.

12. $\dfrac{1}{\sec^2 A} + \dfrac{1}{\csc^2 A} = 1$.

13. $\dfrac{\sec A}{\cos A} - \dfrac{\tan A}{\cot A} = 1$.

14. $\sec A - \tan A \sin A = \cos A$.

15. $\dfrac{\sec A \sin A}{\tan A + \cot A} = \sin^2 A$.

16. $\csc A + \cot A = \dfrac{1 + \cos A}{\sin A}$.

17. $(\sin A + \cos A)(\tan A + \cot A) = \sec A + \csc A$.

18. $(\sin A + \cos A)^2 + (\sin A - \cos A)^2 = 2$.

19. $\dfrac{\sec^2 A}{\sec^2 A - 1} = \csc^2 A$.

20. $\cos^2 A - \sin^2 A = 1 - 2 \sin^2 A$.

21. $\cos^2 A - \sin^2 A = 2 \cos^2 A - 1$.

22. $\tan A - \cot A = \dfrac{1 - 2\cos^2 A}{\sin A \cos A}$.

23. $\tan^2 A \sin^2 A = \tan^2 A - \sin^2 A$.

24. $\dfrac{\sin A}{\csc A} + \dfrac{\cos A}{\sec A} = 1$.

25. $\sin^2 A(1 + \cot^2 A) + \cos^2 A(1 + \tan^2 A) = 2$.

26. $\dfrac{\tan A - 1}{\tan A + 1} = \dfrac{1 - \cot A}{1 + \cot A}$.

27. $\dfrac{1 - \tan^2 x}{1 + \tan^2 x} = 1 - 2\sin^2 x$.

28. $\dfrac{\sin A + \tan A}{1 + \sec A} = \sin A$.

163. Addition theorems for trigonometric functions. If S, A, and B are three *numbers*, then $S(A + B) = SA + SB$ by the distributive law. If we were excessively foolhardy, we might conclude from this that $\sin (A + B) = \sin A + \sin B$; *but this is definitely false.* In fact we should not expect it to be true by analogy, because the symbol sin does not represent a number and the expression $\sin (A + B)$ does not mean the result of multiplying $(A + B)$ by the symbol sin. The falsity of this theorem can be seen from a simple numerical example. For instance, $\sin (30° + 30°) = \sin 60° = \sqrt{3}/2$, while $\sin 30° + \sin 30° = \dfrac{1}{2} + \dfrac{1}{2} = 1$. A true statement is given by the following theorem.

THEOREM 1. *If A and B are any angles,*

$$\sin (A + B) = \sin A \cos B + \cos A \sin B.$$

Proof. Case 1. Suppose the angles A, B, and $A + B$ all terminate in quadrant I. Then we choose a point P on the terminal ray of angle $A + B$ (Fig. 250) and draw PQ perpendicular to the x-axis, PR perpendicular to the terminal ray of angle A, RS perpendicular to PQ and RT perpendicular to the x-axis. Then $\measuredangle RPS = \measuredangle A$ since triangles PRU and QOU are clearly similar.

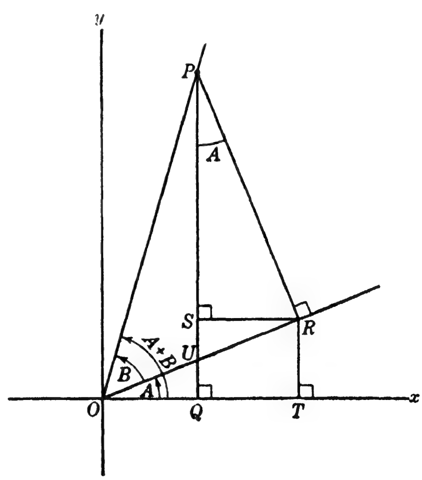


FIG. 250

Now $\sin A \cos B + \cos A \sin B$

$$= \frac{RT}{OR} \cdot \frac{OR}{OP} + \frac{PS}{PR} \cdot \frac{PR}{OP}$$
$$= \frac{RT}{OP} + \frac{PS}{OP} = \frac{RT + PS}{OP}$$
$$= \frac{SQ + PS}{OP} = \frac{PQ}{OP} = \sin (A + B).$$

This proves Case 1.

The remaining cases, where one or more of the angles A, B, $A + B$ terminate in some quadrant other than the first, might be treated similarly keeping in mind that certain functions of angles not in quadrant I may be negative. Other, less tedious devices may be used to prove the theorem for the remaining cases. We shall not complete the proof here.

THEOREM 2. *If A, B are any angles,*

$$\cos (A + B) = \cos A \cos B - \sin A \sin B.$$

Proof. Case 1. Suppose A, B, and $A + B$ terminate in quadrant I. Then (Fig. 250),

$\cos A \cos B - \sin A \sin B$

$$= \frac{OT}{OR} \cdot \frac{OR}{OP} - \frac{SR}{PR} \cdot \frac{PR}{OP}$$
$$= \frac{OT}{OP} - \frac{SR}{OP} = \frac{OT - SR}{OP}$$
$$= \frac{OT - QT}{OP} = \frac{OQ}{OP} = \cos (A + B).$$

This proves Case 1. We shall not complete the remaining cases here.

THEOREM 3. *If A and B are any angles for which all the tangents mentioned exist,*

$$\tan (A + B) = \frac{\tan A + \tan B}{1 - \tan A \tan B}.$$

Proof. $\tan (A + B) = \dfrac{\sin (A + B)}{\cos (A + B)}$

$$= \frac{\sin A \cos B + \cos A \sin B}{\cos A \cos B - \sin A \sin B}$$

(by Theorems 1 and 2)

$$= \frac{\dfrac{\sin A \cos B}{\cos A \cos B} + \dfrac{\cos A \sin B}{\cos A \cos B}}{\dfrac{\cos A \cos B}{\cos A \cos B} - \dfrac{\sin A \sin B}{\cos A \cos B}}$$

(dividing numerator and denominator by $\cos A \cos B$)

$$= \frac{\tan A + \tan B}{1 - \tan A \tan B}.$$

Theorems 1, 2, 3 are called **addition theorems** for the sine, cosine, and tangent, respectively.

THEOREM 4. *If A and B are any angles,*

$$\sin (A - B) = \sin A \cos B - \cos A \sin B.$$

Proof. $\sin (A - B) = \sin (A + [-B])$

(1) $\qquad = \sin A \cos [-B] + \cos A \sin [-B].$

But $\cos [-B] = \cos B$ and $\sin [-B] = -\sin B$. Making these substitutions in (1) we have the theorem.

THEOREM 5. *If A and B are any angles,*

$$\cos (A - B) = \cos A \cos B + \sin A \sin B.$$

Proof.
$$\begin{aligned} \cos (A - B) &= \cos (A + [-B]) \\ &= \cos A \cos [-B] - \sin A \sin [-B] \\ &= \cos A \cos B + \sin A \sin B. \end{aligned}$$

THEOREM 6. *If A and B are any angles for which all the tangents mentioned exist,*

$$\tan (A - B) = \frac{\tan A - \tan B}{1 + \tan A \tan B}.$$

Proof.
$$\begin{aligned} \tan (A - B) &= \tan (A + [-B]) \\ &= \frac{\tan A + \tan [-B]}{1 - \tan A \tan [-B]} \end{aligned}$$

But $\tan [-B] = -\tan B$. Making this substitution, we have the theorem.

Remark. Theorems 4, 5, 6 are easy to remember since each may be obtained in turn from theorems 1, 2, 3, respectively, merely by changing all the *visibly* written signs.

Example 1. Find, without tables, the value of sin 75°.

$$\sin 75^\circ = \sin(45^\circ + 30^\circ) = \sin 45^\circ \cdot \cos 30^\circ + \cos 45^\circ \cdot \sin 30^\circ = \frac{\sqrt{2}}{2}\cdot\frac{\sqrt{3}}{2} + \frac{\sqrt{2}}{2}\cdot\frac{1}{2}$$
$$= \tfrac{1}{4}(\sqrt{6} + \sqrt{2}).$$

Example 2. Find, without tables, the value of sin 15°.

$$\sin 15^\circ = \sin(45^\circ - 30^\circ) = \sin 45^\circ \cdot \cos 30^\circ - \cos 45^\circ \cdot \sin 30^\circ = \frac{\sqrt{2}}{2}\cdot\frac{\sqrt{3}}{2} - \frac{\sqrt{2}}{2}\cdot\frac{1}{2}$$
$$= \tfrac{1}{4}(\sqrt{6} - \sqrt{2}).$$

Example 3. Express $\sin(A + 90^\circ)$, $\cos(A + 90^\circ)$, $\tan(A + 90^\circ)$ in terms of functions of A alone.

$$\begin{aligned}\sin(A + 90^\circ) &= \sin A \cdot \cos 90^\circ + \cos A \cdot \sin 90^\circ\\ &= (\sin A)\cdot 0 + (\cos A)\cdot 1 = \cos A,\\ \cos(A + 90^\circ) &= \cos A \cdot \cos 90^\circ - \sin A \cdot \sin 90^\circ\\ &= -\sin A,\\ \tan(A + 90^\circ) &= \frac{\sin(A + 90^\circ)}{\cos(A + 90^\circ)} = \frac{\cos A}{-\sin A} = -\cot A.\end{aligned}$$

Why can we not use theorem 3 directly to obtain the last formula?

EXERCISES

1. Show that $\cos 15^\circ = \frac{1}{4}(\sqrt{6} + \sqrt{2}) = \sin 75^\circ$.

2. Show that $\cos 75^\circ = \frac{1}{4}(\sqrt{6} - \sqrt{2}) = \sin 15^\circ$.

3. Show that $\tan 75^\circ = 2 + \sqrt{3} = \cot 15^\circ$.

4. Show that $\tan 15^\circ = 2 - \sqrt{3} = \cot 75^\circ$.

If A and B are both in the first quadrant and if $\sin A = \frac{3}{5}$ *and* $\cos B = \frac{5}{13}$, *find:*

5. $\sin(A + B)$. **6.** $\cos(A + B)$. **7.** $\tan(A + B)$.

8. $\sin(A - B)$. **9.** $\cos(A - B)$. **10.** $\tan(A - B)$.

If A and B are both in the first quadrant and if $\tan A = \frac{4}{3}$ *and* $\sec B = \frac{13}{12}$, *find:*

11. $\sin(A + B)$. **12.** $\cos(A + B)$. **13.** $\tan(A + B)$.

14. $\sin(A - B)$. **15.** $\cos(A - B)$. **16.** $\tan(A - B)$.

If neither A nor B are in the first quadrant and if $\sin A = \frac{4}{5}$ *and* $\cos B = \frac{12}{13}$, *find:*

17. $\sin (A + B)$. **18.** $\cos (A + B)$. **19.** $\tan (A + B)$.

20. $\sin (A - B)$. **21.** $\cos (A - B)$. **22.** $\tan (A - B)$.

If neither A nor B are in the first quadrant and if $\tan A = \frac{3}{4}$ *and* $\sec B = \frac{13}{5}$, *find:*

23. $\sin (A + B)$. **24.** $\cos (A + B)$. **25.** $\tan (A + B)$.

26. $\sin (A - B)$. **27.** $\cos (A - B)$. **28.** $\tan (A - B)$.

29. Prove that $\sin (A + 45°) = \frac{\sqrt{2}}{2} (\sin A + \cos A)$.

30. Prove that $\cos (A + 45°) = \frac{\sqrt{2}}{2} (\cos A - \sin A)$.

31. Prove that $\sin (45° + A) = \cos (45° - A)$.

32. From the functions of 30° and 180° find the functions of (*a*) 210°; (*b*) 150°.

33. From the addition theorems, deduce the formulas for $\sin (A + 180°)$, $\cos (A + 180°)$, $\tan (A + 180°)$.

34. From theorems 4, 5, and 6, deduce the formulas for $\sin (180° - A)$, $\cos (180° - A)$, $\tan (180° - A)$.

35. From theorems 4 and 5 deduce the formulas for $\sin (90° - A)$, $\cos (90° - A)$.

36. From theorems 4, 5, and 6, deduce the formulas for $\sin (-A)$, $\cos (-A)$, $\tan (-A)$. (Hint: Let $-A = 0° - A$.)

37. From theorems 4, 5, and 6, deduce the formulas for $\sin (A - 180°)$, $\cos (A - 180°)$, $\tan (A - 180°)$.

Verify the following identities:

38. $\dfrac{\sin (A + B)}{\sin (A - B)} = \dfrac{\tan A + \tan B}{\tan A - \tan B}$.

39. $\sin (A + B) + \sin (A - B) = 2 \sin A \cos B$.

40. $\sin (A + B) - \sin (A - B) = 2 \cos A \sin B$.

41. $\cos (A + B) + \cos (A - B) = 2 \cos A \cos B$.

42. $\cos (A + B) - \cos (A - B) = -2 \sin A \sin B$.

43. Rewrite the formulas of exercises 39 – 42, in terms of x and y alone, letting $A + B = x$ and $A - B = y$ and therefore (why?) $A = \frac{x + y}{2}$ and $B = \frac{x - y}{2}$.

164. Double and half angle formulas. If in the identity $\sin (A + B) = \sin A \cos B + \cos A \sin B$, we take $B = A$ then $A + B$ becomes $A + A$ or $2A$ and we obtain

$$(1) \qquad \sin 2A = 2 \sin A \cos A.$$

Making the same substitution $B = A$ in the formulas for $\cos (A + B)$ and $\tan (A + B)$ respectively, we obtain

(2) $$\cos 2A = \cos^2 A - \sin^2 A,$$
and

(3) $$\tan 2A = \frac{2\tan A}{1 - \tan^2 A}.$$

Formula (2) may be rewritten as

(4) $$\cos 2A = 1 - 2\sin^2 A$$

or

(5) $$\cos 2A = 2\cos^2 A - 1$$

by merely replacing $\cos^2 A$ by $1 - \sin^2 A$, or $\sin^2 A$ by $1 - \cos^2 A$, respectively, in (2), and simplifying.

Remark. Note that $\sin 2A$ is *not* $2 \sin A$, and so on. One should not expect any analogy with the commutative law for multiplication (since we are *not* multiplying by the word sin) any more than one should expect the commutative law to operate in spelling. If it *were* true that $\sin nx = n \sin x$ (it is not!), then we would need no table of sines. For, to find the sine of any number of degrees, say 38°, we would have only to multiply the sine of 1° by 38. Thus we would need only the value of sin 1°.

Denote the angle $2A$ by x. Then $\frac{x}{2} = A$.

From (4) we obtain

$$\cos x = 1 - 2\sin^2\frac{x}{2}.$$

Hence,

$$2\sin^2\frac{x}{2} = 1 - \cos x,$$

$$\sin^2\frac{x}{2} = \frac{1 - \cos x}{2},$$

(6) $$\sin\frac{x}{2} = \pm\sqrt{\frac{1 - \cos x}{2}}.$$

From (5) we obtain, similarly,

$$\cos x = 2\cos^2\frac{x}{2} - 1,$$

$$2\cos^2\frac{x}{2} = 1 + \cos x,$$

$$\cos^2\frac{x}{2} = \frac{1 + \cos x}{2},$$

(7) $$\cos\frac{x}{2} = \pm\sqrt{\frac{1+\cos x}{2}}.$$

The proper sign in (6) and (7) may be chosen by determining the quadrant in which $\frac{x}{2}$ terminates in any given example.

Combining (6) and (7) we obtain

$$\tan\frac{x}{2} = \frac{\sin\frac{x}{2}}{\cos\frac{x}{2}} = \frac{\pm\sqrt{\frac{1-\cos x}{2}}}{\pm\sqrt{\frac{1+\cos x}{2}}}$$
$$= \pm\sqrt{\frac{1-\cos x}{2}\cdot\frac{2}{1+\cos x}}$$

or

(8) $$\tan\frac{x}{2} = \pm\sqrt{\frac{1-\cos x}{1+\cos x}}.$$

The formulas

(9) $$\tan\frac{x}{2} = \frac{\sin x}{1+\cos x},$$

and

(10) $$\tan\frac{x}{2} = \frac{1-\cos x}{\sin x},$$

can be obtained from exercises 22 and 23 below, by replacing $2A$ by x and hence A by $x/2$.

Example. Without tables find $\sin 22\frac{1}{2}°$, $\cos 22\frac{1}{2}°$, $\tan 22\frac{1}{2}°$. If we take $x = 45°$, we have $\frac{x}{2} = 22\frac{1}{2}°$. Hence, from (6),

$$\sin 22\tfrac{1}{2}° = \pm\sqrt{\frac{1-\cos 45°}{2}}$$
$$= \pm\sqrt{\frac{1-\frac{\sqrt{2}}{2}}{2}} = \pm\sqrt{\frac{2-\sqrt{2}}{4}}$$
$$= \pm\tfrac{1}{2}\sqrt{2-\sqrt{2}}.$$

But $22\frac{1}{2}°$ is in quadrant I; hence its sine is positive. Therefore $\sin 22\frac{1}{2}° = \frac{1}{2}\sqrt{2-\sqrt{2}}$. Similarly, from (7), $\cos 22\frac{1}{2}° = \frac{1}{2}\sqrt{2+\sqrt{2}}$.

From (8), $\tan 22\frac{1}{2}° = \sqrt{\dfrac{1 - \dfrac{\sqrt{2}}{2}}{1 + \dfrac{\sqrt{2}}{2}}} = \sqrt{\dfrac{2 - \sqrt{2}}{2 + \sqrt{2}}}.$

From (9), $\tan 22\frac{1}{2}° = \dfrac{\dfrac{\sqrt{2}}{2}}{1 + \dfrac{\sqrt{2}}{2}} = \dfrac{\sqrt{2}}{2 + \sqrt{2}} = \dfrac{1}{\sqrt{2} + 1}.$

From (10), $\tan 22\frac{1}{2}° = \dfrac{1 - \dfrac{\sqrt{2}}{2}}{\dfrac{\sqrt{2}}{2}} = \dfrac{2 - \sqrt{2}}{\sqrt{2}}$

$$= \frac{2\sqrt{2} - 2}{2} = \sqrt{2} - 1.$$

EXERCISES

1. Find, without tables, sin 15°, cos 15°, tan 15°.

2. From the known functions of 150° find sin 75°, cos 75°, tan 75° without tables.

If $\sin A = \frac{4}{5}$ *and* A *is in the first quadrant, find:*

3. $\sin 2A$. **4.** $\cos 2A$. **5.** $\tan 2A$.

6. $\sin \frac{A}{2}$. **7.** $\cos \frac{A}{2}$. **8.** $\tan \frac{A}{2}$.

If $\sin A = \frac{4}{5}$ *and* A *is not in the first quadrant, find:*

9. $\sin 2A$. **10.** $\cos 2A$. **11.** $\tan 2A$.

12. $\sin \frac{A}{2}$. **13.** $\cos \frac{A}{2}$. **14.** $\tan \frac{A}{2}$.

15. Show that $\sin 3x = 3 \sin x - 4 \sin^3 x$. (Hint: Let $A = x$ and $B = 2x$ in the formula for $\sin (A + B)$; then use the formula for $\sin 2x$ and simplify.)

16. Show that $\cos 3x = 4 \cos^3 x - 3 \cos x$. (See hint for exercise 15.)

17. From the known functions of 135° find, without tables, sin $67\frac{1}{2}°$, cos $67\frac{1}{2}°$, tan $67\frac{1}{2}°$.

18. Rewrite the formulas (1)–(5) in terms of x letting $2A = x$ and therefore $A = \frac{x}{2}$.

Verify the following identities:

19. $\cos 2x = \dfrac{1 - \tan^2 x}{1 + \tan^2 x}$.

20. $\sin 2x = \dfrac{2 \tan x}{1 + \tan^2 x}$.

21. $4 \csc^2 2x = \sec^2 x(1 + \cot^2 x)$.

22. $\dfrac{\sin 2A}{1 + \cos 2A} = \tan A$.

23. $\dfrac{1 - \cos 2A}{\sin 2A} = \tan A$.

24. $\dfrac{1 + \cos 2A}{\sin 2A} = \cot A$.

25. $\sec 2A = \dfrac{1}{1 - 2\sin^2 A}$.

26. $\dfrac{1 - \tan A}{1 + \tan A} = \dfrac{1 - \sin 2A}{\cos 2A}$.

165. Inverse functions. The symbol **arc sin x** means the angle whose sine is x. Since many angles may have the same sine, arc sin x is not a single-valued function of x. For example, arc sin $\frac{1}{2}$ = 30°, 150° or any angle obtained from either one of these by adding an integral multiple of 360°, since any one of these angles has $\frac{1}{2}$ as its sine. These facts are written in the following concise form:

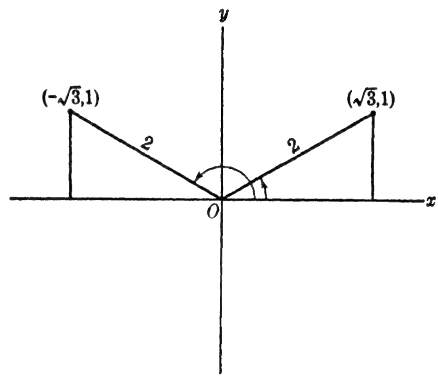


FIG. 251

$$\text{(1)} \qquad \text{arc sin } \tfrac{1}{2} = \begin{cases} 30° + n \cdot 360° \\ 150° + n \cdot 360° \end{cases}$$ where n is any integer, positive, negative or zero.

If we want to find those angles whose sine is $\frac{1}{2}$ which lie in a given range, such as between − 360° and + 360°, we let n take various integral values, as follows: for $n = 0$, we get 30° and 150°; for $n = -1$, we get − 330° and − 210°; for other integral values of n we get values outside the prescribed range. Hence, 30°, 150°, − 330°, and − 210° are the solutions desired. The right member of (1) may be called the **general value** of arc sin $\frac{1}{2}$.

Similarly, $\text{arc cos } \frac{1}{2} = \begin{cases} 60° + n \cdot 360° \\ 300° + n \cdot 360° \end{cases}$

where n is any integer. The particular values lying between + 180° and − 180° are 60° and − 60°.

Example. $\text{arc tan } 1 = \begin{cases} 45° + n\, 360° \\ 225° + n\, 360° \end{cases}$ where n is any integer.

The particular values lying between $+ 180°$ and $- 180°$ are $45°$ and $- 135°$.

Remark. arc sin x is also written as $\sin^{-1}x$. Note that the $- 1$ here is not meant to be an exponent. This symbol should not be confused with $(\sin x)^{-1}$ which is the symbol always used for sin x raised to the exponent $- 1$. Similarly $\cos^{-1}x$ and $\tan^{-1}x$ are often written for arc cos x and arc tan x respectively.

The functions arc sin x, arc cos x, etc., are called **inverse trigonometric functions.** If $y = \sin x$ then $x = \text{arc sin } y$, and so on.

To avoid the confusion engendered by the many-valuedness of these symbols, we sometimes restrict ourselves to the so-called principal values. The **principal value** of an inverse trigonometric function is the value with the smallest absolute value; if two values have equal absolute values we select the positive value.

Example. The principal value of arc sin $(1/2)$ is $30°$. The principal value of arc sin $(- 1/2)$ is $- 30°$. The principal value of arc cos $(1/2)$ is $60°$, not $- 60°$. The principal value of arc tan $(- 1)$ is $- 45°$.

EXERCISES

Give the general value, the particular values lying between $- 360°$ and $+ 360°$, and the principal value of each of the following:

1. arc sin $(1/2)$.
2. arc sin $(- 1/2)$.
3. arc cos $(1/2)$.
4. arc cos $(- 1/2)$.
5. arc tan 1.
6. arc tan $(- 1)$.
7. arc sin $(\sqrt{3}/2)$.
8. arc cos $(-\sqrt{3}/2)$.
9. arc tan $\sqrt{3}$.
10. arc sin $(\sqrt{2}/2)$.
11. arc cos $(-\sqrt{2}/2)$.
12. arc sin $(-\sqrt{2}/2)$.
13. arc cos $(\sqrt{2}/2)$.

Using principal values, evaluate the following expressions:

14. sin (arc sin $(1/3)$).
15. cos (arc cos $(3/4)$).
16. tan (arc tan $(4/5)$).
17. sin (arc cos $(1/3)$).
18. cos (arc sin $(3/4)$).
19. tan (arc cot $(4/5)$).
20. sin (arc tan 2).
21. cos (arc sec 2).
22. tan (arc cos $(4/5)$).

166. Trigonometric equations. The distinction between identical and conditional equations (section 32) holds just as well among equations involving trigonometric functions. By a **trigonometric equation** we shall mean a *conditional* equation involv-

ing trigonometric functions (as opposed to identities). While polynomial equations always have a finite number of roots, the simplest trigonometric equations have an infinite number of roots. For example, the equation $\sin x = 1/2$ has the *general solution* $x = \text{arc sin } (1/2) = \begin{cases} 30° + n\,360° \\ 150° + n\,360° \end{cases}$ where n is any integer. We may wish, at times, to pick out the *particular solutions* lying in a certain range. The particular solutions lying between $-360°$ and $+360°$ in this example are $30°$, $150°$, $-210°$, $-330°$.

Example 1. The equation $2 \cos x - 1 = 0$ is solved by first solving for $\cos x = 1/2$. Then $x = \text{arc cos } (1/2)$. The student should give the general solution and the particular solutions lying between $-360°$ and $+360°$.

It is usually best to solve for the values of some trigonometric function and then for the angle.

Example 2. Give the general solution and the particular solutions lying between $+360°$ and $-360°$ of the equation

$$2 \sin^2 x - \sin x - 1 = 0.$$

Factoring we get

$$(2 \sin x + 1)(\sin x - 1) = 0.$$

Therefore either

$$2 \sin x + 1 = 0 \quad \text{or} \quad \sin x - 1 = 0.$$

Or,

$$\sin x = -1/2 \quad \text{or} \quad \sin x = 1.$$

Hence

$$x = \text{arc sin } (-1/2) \quad \text{or} \quad x = \text{arc sin } 1.$$

Or, $$x = \begin{cases} -30° + n \cdot 360° \\ -150° + n \cdot 360° \end{cases} \quad \text{or} \quad x = 90° + n \cdot 360°.$$

This is the general solution. The particular solutions in the desired range are $-30°$, $-150°$, $90°$, $330°$, $210°$, $-270°$.

Example 3. Give the general solution and the particular solutions lying between $-360°$ and $+360°$ inclusive of the equation $\sin 2x + \sin x = 0$. This becomes

$$2 \sin x \cos x + \sin x = 0,$$

$$\sin x(2 \cos x + 1) = 0,$$

$$\sin x = 0, \quad \text{or} \quad 2\cos x + 1 = 0,$$
$$\cos x = -1/2$$

$$x = \text{arc} \sin 0, \qquad x = \text{arc} \cos(-1/2)$$

$$x = \begin{cases} 0° + n \cdot 360°, \\ 180° + n \cdot 360° \end{cases} \qquad x = \begin{cases} 120° + n \cdot 360° \\ 240° + n \cdot 360°. \end{cases}$$

This is the general solution. The desired particular solutions are 0°, 360°, − 360°, 180°, − 180°, 120°, − 240°, 240°, − 120°.

EXERCISES

Give the general solution and the particular solutions lying between − 360° and + 360° inclusive:

1. $2\sin x - 1 = 0$. **2.** $4\sin^2 x = 3$. **3.** $4\cos^2 x - 3 = 0$.
4. $\tan^2 x - 3 = 0$. **5.** $2\sin^2 x + \sin x - 1 = 0$.
6. $\csc^2 x - 2 = 0$. **7.** $2\cos^2 x - \cos x - 1 = 0$.
8. $\sin^2 x = \sin x$. **9.** $2\cos x \sin x - \cos x = 0$.
10. $\sin 2x = 2\sin x$. **11.** $\sin 2x = \cos x$. **12.** $\cos 2x = \sin x$.
13. $\sin 2x = -\sin x$. **14.** $\cos 2x = -\cos x$.
15. $\cos 2x + \sin x - 1 = 0$. **16.** $\cos 2x - \cos x + 1 = 0$.
17. $\cos 2x = \cos^2 x$. **18.** $\cos 2x = -\sin^2 x$. **19.** $\cos 2x + \sin x + 1 = 0$.
20. $\sin 2x = 2\cos^2 x$. **21.** $\sin 2x = 1/2$. **22.** $\sin 2x = \sqrt{2}/2$.

167. The ambiguous case. In section 127, we pointed out that we might expect to be able to solve a triangle when enough data are specified so that all triangles having the given measurements must be congruent to each other. In this section we shall point out that the law of sines enables us to accomplish the solution of triangles in a case that was beyond our expectations, namely when two sides and the angle opposite one of them are given. That "side, side, angle" are insufficient to fix the size and shape of the triangle may be seen from an attempt to construct the triangle from the data. Suppose A, b, a are given. We may first construct angle A and, on one side of it lay off side b (Fig. 252). Then we swing an arc with center C and radius a. There are several possibilities. If a is too short to hit the base line there is no triangle having the given parts (Fig. 252a). If a happens to be exactly of the right length to touch the base line just once (the arc is then tangent to the base line) then there is

just one triangle having the given parts and it must be a right triangle (Fig. 252b). If a is longer than the radius of this tangent arc but shorter than b, then there are two different (non-congruent) triangles ABC and $AB'C$ each of which has the given parts (Fig. 252c). If a is longer than b, the swinging arc hits the base line twice but one of the triangles ($AB'C$) is unsatisfactory

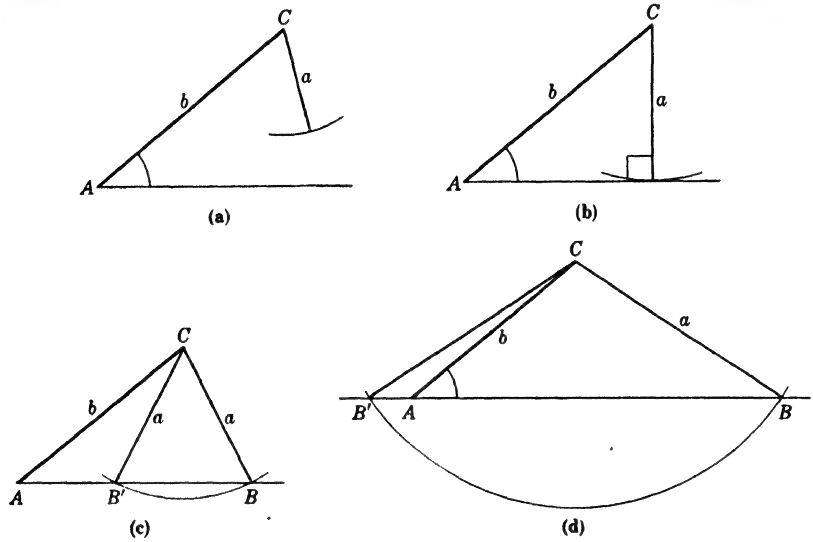


FIG. 252

(Fig. 252d) since it does not contain angle A; if B' coincides with A, this also yields just one triangle ABC. In case (b), $a = b \sin A$. In case (a), $a < b \sin A$. In case (c), $b \sin A < a < b$. In case (d), $a \geqq b$. By testing to see which of these conditions holds, we can tell in advance which situation occurs. For obvious reasons, the case where "side, side, angle" are given is called the **ambiguous case.** In case (c) there are two satisfactory triangles, but, fortunately, the law of sines will enable us to obtain both of them, as in the following example.

Example. If $A = 30°$, $b = 10$, $a = 8$, solve the triangle (or triangles). Now $b \sin A = 10(1/2) = 5$. Hence $b \sin A < a < b$ since $5 < 8 < 10$. Therefore there are two solutions, as in Fig. 252c. By the law of sines,

$$\frac{\sin B}{10} = \frac{\sin 30°}{8},$$

$$\sin B = \frac{10(1/2)}{8} = \frac{5}{8} = .625.$$

Looking in the tables, one angle whose sine is .625 is 39° (approx.). But we must recall that sin $(180° - B) = \sin B$. Hence $180° - 39° = 141°$ is another angle whose sine is .625. In Fig. 252c, if $\measuredangle ABC = 39°$ then $\measuredangle AB'C = 141°$. This gives us the necessary information to solve each of the triangles ABC and $AB'C$ separately since, in each one, we now know all the angles (why?) and two of the sides. In each case the remaining side may be found as usual by the law of sines. The student should complete the solution of both triangles.

EXERCISES

Solve each of the following triangles, given that:

1. $A = 35°$, $b = 10$, $a = 8$.

2. $B = 40°$, $c = 20$, $b = 16$.

3. $A = 30°$, $b = 10$, $a = 4$.

4. $C = 30°$, $b = 20$, $c = 10$.

5. $A = 30°$, $b = 10$, $a = 12$.

6. $B = 45°$, $a = 50$, $b = 40$.

7. $A = 30°$, $b = 10$, $a = 10$.

8. $C = 48°$, $a = 100$, $c = 85$.

9. The distance of the earth E from the sun S is approximately 93 million miles. The distance of Venus from the sun S is approximately 67 million miles. If the angle between Venus and the sun (at the earth) is observed at a certain time to be 20°, how far from the earth is Venus at that time? (Two answers. See Fig. 253.)

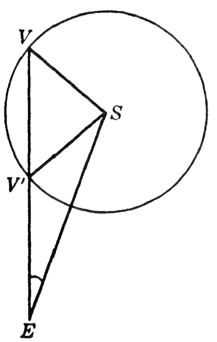


FIG. 253

168. Completing the square. The expansion of the perfect square $(x + h)^2$ is $x^2 + 2xh + h^2$. The coefficient of the first power of x is $2h$. Clearly the constant term may be obtained by taking half of this coefficient of x and squaring it.

Example 1. Solve the equation $x^2 + 6x + 4 = 0$ by completing the square.

We write $x^2 + 6x = -4$. To create a perfect square on the left we take half of 6 or 3 and add its square $3^2 = 9$ to both sides, obtaining $x^2 + 6x + 9 = 9 - 4$ or $x^2 + 6x + 9 = 5$. This can be written as $(x + 3)^2 = 5$. Hence $x + 3 = \pm\sqrt{5}$ or $x = -3 \pm \sqrt{5}$.

Example 2. Find the center and radius of the circle whose equation is $x^2 - 4x + y^2 - 6y = 3$.

To complete the square for the x terms we must add 4 to both sides and to complete the square for the y terms we must add 9 to both sides. Hence, we get

$$x^2 - 4x + 4 + y^2 - 6y + 9 = 3 + 4 + 9$$

or

$$(x - 2)^2 + (y - 3)^2 = 16.$$

Therefore the center is (2,3) and the radius is 4.

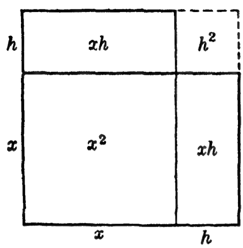


FIG. 254

Example 3. To derive the formula for the roots of the general quadratic equation $ax^2 + bx + c = 0$ by completing the square we write

$$ax^2 + bx = - c,$$
$$x^2 + \frac{b}{a}x = - \frac{c}{a}.$$

Half of the coefficient of x is $\frac{b}{2a}$. Therefore we add $\left(\frac{b}{2a}\right)^2$ or $\frac{b^2}{4a^2}$ to both sides, obtaining

$$x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} = \frac{b^2}{4a^2} - \frac{c}{a}$$

or

$$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}.$$

The remainder of the proof is left to the student.

The geometric significance of the phrase "completing the square" is obvious from Fig. 254.

EXERCISES

(*a*) *Solve each of the equations in the exercises on page 141 by completing the square.*

(*b*) *Find the center and radius of each of the following circles:*

1. $x^2 + y^2 - 6x + 8y = 0.$
2. $x^2 + y^2 - 6x + 4y = 12.$
3. $x^2 - 2x + y^2 + 6y = 6.$
4. $x^2 - 4x + y^2 + 10y = - 20.$

169. The remainder and factor theorems. When 163 is divided by 12 we obtain a quotient of 13 and a remainder of 7. The work is sometimes arranged as follows:

$$\begin{array}{r} 13 \\ 12\overline{)163} \\ \underline{12} \\ 43 \\ \underline{36} \\ 7 \end{array}$$

This result is expressed as $\dfrac{163}{12} = 13 + \dfrac{7}{12}$, or, more simply, as $163 = 12 \cdot 13 + 7$. We would not consider the division accomplished if we wrote $163 = 12 \cdot 10 + 43$ or $163 = 12 \cdot 12 + 19$ because, although the latter statements are true, they exhibit a proposed remainder which is not less than the divisor 12 and such a remainder permits the division to be carried further. This suggests the following definition.

DEFINITION 1. *To* ***divide*** *a natural number f by a natural number d, called the divisor, means to find natural numbers q and r, called the quotient and remainder respectively, such that*

$$f = dq + r$$

where r is less than d. If $r = 0$, d is said to be a ***factor*** *of f.*

Note that $f - dq = r$. Division of natural numbers is essentially repeated subtraction just as multiplication of natural numbers is repeated addition. For example, as seen in the familiar scheme for division given above, we subtract thirteen twelves from 163 to get a remainder of 7.

Analogously we make the following definition for polynomials.

DEFINITION 2. *To* ***divide*** *a polynomial $f(x)$ by a polynomial $d(x)$, called the divisor, means to find polynomials $q(x)$ and $r(x)$, called the quotient and remainder respectively, such that*

$$f(x) = d(x) \cdot q(x) + r(x)$$

where the degree of $r(x)$ is less than the degree of $d(x)$. If $r(x)$ is zero, then $d(x)$ is said to be a ***factor*** *of $f(x)$.*

For example, dividing $2x^3 + 5x^2 + 7x + 4$ by $x^2 + x + 1$ we proceed as follows:

$$\begin{array}{r|l}
 & 2x + 3 \\
\cline{2-2}
x^2 + x + 1 & 2x^3 + 5x^2 + 7x + 4 \\
 & 2x^3 + 2x^2 + 2x \\
\cline{2-2}
 & \qquad 3x^2 + 5x + 4 \\
 & \qquad 3x^2 + 3x + 3 \\
\cline{2-2}
 & \qquad\qquad 2x + 1
\end{array}$$

obtaining the quotient $2x + 3$ and remainder $2x + 1$. The result may be expressed as

$$\frac{2x^3 + 5x^2 + 7x + 4}{x^2 + x + 1} = 2x + 3 + \frac{2x + 1}{x^2 + x + 1}$$

or more simply as

$$2x^3 + 5x^2 + 7x + 4 = (x^2 + x + 1)(2x + 3) + (2x + 1).$$

We would not consider the division accomplished if we wrote

$$2x^3 + 5x^2 + 7x + 4 = (x^2 + x + 1) \cdot 2x + (3x^2 + 5x + 4)$$

although this is a true statement, because the proposed remainder $3x^2 + 5x + 4$ does not have lower degree than the divisor and therefore permits the division to be carried further.

If we divide by a divisor of degree one, the remainder must be a constant (that is, a polynomial of degree "zero") since its degree must be less than the degree of the divisor, and therefore it can not involve x at all. Hence to divide a polynomial $f(x)$ by a divisor of the form $x - a$, where a is a constant, means to find a quotient polynomial $q(x)$ and a constant remainder r such that

$$f(x) = (x - a)q(x) + r. \tag{1}$$

For example, dividing $f(x) = x^2 - 5x + 10$ by $d(x) = x - 2$ we obtain a quotient of $q(x) = x - 3$ and a remainder of $r = 4$. We may write this as $x^2 - 5x + 10 = (x - 2)(x - 3) + 4$. Note that $f(2) = 2^2 - 5 \cdot 2 + 10 = 4$ which is precisely the remainder. That this is no accident is seen from the following theorem.

Remainder Theorem. *If a polynomial $f(x)$ is divided by a divisor of the form $x - a$ the constant remainder r is equal to $f(a)$.*

Proof. The division may be expressed by (1). But (1) is an identity; that is, it is true for all values of x. In particular it is true for $x = a$. Substituting $x = a$ in (1) we obtain

$$f(a) = (a - a) \cdot q(a) + r$$

or

$$f(a) = r.$$

This proves the theorem.

FACTOR THEOREM. *If* $(x - a)$ *is a factor of the polynomial* $f(x)$ *then* a *is a root of the equation* $f(x) = 0$, *and conversely.*

Proof. If $(x - a)$ is a factor of $f(x)$ then the remainder obtained by dividing $f(x)$ by $(x - a)$ is zero. But by the previous theorem, this remainder is $f(a)$. Hence $f(a) = 0$, which means that a is a root of the equation $f(x) = 0$. This proves the theorem.

To prove the converse, we observe that if $f(a) = 0$ then $r = 0$ and $f(x) = (x - a)q(x)$.

Remark. It follows from this that every equation $f(x) = 0$ of degree n has n roots, provided a root r is counted as many times as $x - r$ occurs as a factor of $f(x)$. (For example, the equation $x^3 - 6x^2 + 12x - 8 = 0$ or $(x - 2)(x - 2)(x - 2) = 0$ has only one root, 2, but we would count it as three roots because there are three factors $x - 2$.) For by the fundamental theorem of algebra (section 37) the equation $f(x) = 0$ has at least one root r_1. By the factor theorem $(x - r_1)$ is a factor of $f(x)$. That is $f(x) = (x - r_1)f_1(x)$ where $f_1(x)$ is a polynomial of degree $n - 1$. But $f_1(x) = 0$ has a root r_2 by the fundamental theorem of algebra. Hence $(x - r_2)$ is a factor of $f_1(x)$, or $f_1(x) = (x - r_2)f_2(x)$ where $f_2(x)$ is of degree $n - 2$. That is, $f(x) = (x - r_1)(x - r_2)$ $f_2(x)$. Clearly this process can be continued to yield n linear factors,

$$f(x) = (x - r_1)(x - r_2)\cdots(x - r_n) \cdot f_n$$

where f_n is a constant (degree $n - n = 0$). It can also be proved that a polynomial can be thus factored in only one way, apart from the order in which the factors are written down. Therefore, $r_1, r_2, \cdots, r_n$ are the n roots of $f(x) = 0$. Some or all of them may be equal. A root r is called a root of **multiplicity** k if $(x - r)$ occurs as a factor of $f(x)$ exactly k times. Roots of multiplicity 2 or 3 are called **double** or **triple** roots, respectively.

EXERCISES

Divide by long division until a constant remainder is obtained and then find the remainder again by means of the remainder theorem:

1. $(x^2 - 5x + 7) \div (x - 3)$.
2. $(3x^2 - x - 9) \div (x - 4)$.
3. $(x^3 - 2x^2 + x - 3) \div (x - 2)$.
4. $(x^3 - 5) \div (x - 1)$.
5. $(x^2 - 5x + 6) \div (x + 2)$.
6. $(x^2 - 5x + 6) \div (x - 2)$.
7. $(x^3 - 2x^2 + x - 3) \div (x + 2)$.
8. $(x^3 - 5) \div (x + 1)$.

By means of the factor theorem, decide whether or not each of the following statements is true:

9. $(x - 2)$ is a factor of $x^3 - x^2 - 5x + 6$.
10. $(x - 3)$ is a factor of $2x^3 - 6x^2 - 5x + 15$.
11. $(x + 1)$ is a factor of $x^3 + 3x^2 + 3x + 1$.
12. $(x + 1)$ is a factor of $x^3 + x^2 - x - 1$.
13. $(x - 1)$ is a factor of $2x^3 - x^2 - 4x + 5$.
14. $(x - 1)$ is a factor of $2x^3 - 3x^2 + 4x - 5$.
15. $(x + 2)$ is a factor of $3x^3 + 5x^2 - 7x - 10$.
16. $(x + 3)$ is a factor of $x^3 + 27$.
17. $(x - 3)$ is a factor of $x^3 + 27$.
18. $x - y$ is a factor of $x^n - y^n$ if n is any natural number.
19. $x - y$ is a factor of $x^n + y^n$ if n is even.
20. $x + y$ is a factor of $x^n + y^n$ if n is odd.
21. $x + y$ is a factor of $x^n - y^n$ if n is even.

170. Rational roots of a polynomial equation with integral coefficients. We might try to find the rational roots of an equation by trial and error; that is, by substituting all possible rational numbers to see if they satisfy the equation. But since there are infinitely many rational numbers (section 24 or 147) this is a gloomy prospect. The theorem below will enable us to cut the number of trials down to a finite number, provided the coefficients of the equation are integers. We shall need the following preliminary theorem, which we shall not prove.

If A, B, C are integers and C is a factor of AB while C has no factor, except ± 1, in common with A then C must be a factor of B.

Examples. Let $A = 5$, $B = 12$, $C = 6$. C has no factor except ± 1 in common with A, while C is a factor of $AB = 60$; therefore C must be a factor of $B = 12$. On the other hand if $A = 4$, $B = 9$, $C = 6$ then C is a factor of $AB = 36$ while C is

not a factor of either A or B; the theorem does not apply here because C has a factor 2 in common with A and a factor 3 in common with B.

THEOREM. *If* $f(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n = 0$ *has integral coefficients and if* p/q *is a rational root, reduced to lowest terms, then* p *is a factor of* a_n *and* q *is a factor of* a_0.

Proof. By hypothesis

$$a_0\frac{p^n}{q^n} + a_1\frac{p^{n-1}}{q^{n-1}} + a_2\frac{p^{n-2}}{q^{n-2}} + \cdots + a_{n-1}\frac{p}{q} + a_n = 0.$$

Multiplying by q^n we have

$$(1) \quad a_0p^n + a_1p^{n-1}q + a_2p^{n-2}q^2 + \cdots + a_{n-1}pq^{n-1} + a_nq^n = 0.$$

Transposing the last term and factoring we have

$$p(a_0p^{n-1} + a_1p^{n-2}q + a_2p^{n-3}q^2 + \cdots + a_{n-1}q^{n-1}) = -a_nq^n.$$

Since the a's and p and q are all integers and since sums and products of integers are again integers, the parenthesis represents an integer. Hence p is a factor of the integer on the left of the equation. But therefore p is a factor of the equal integer on the right. Since p/q is reduced to lowest terms, p has no factor except ± 1 in common with q and hence * no factor except ± 1 in common with q^n. Let $A = -q^n$, $B = a_n$ and $C = p$ in the preliminary theorem above. Since p is a factor of $-a_nq^n$ and has no factor (except ± 1) in common with $-q^n$, p must be a factor of a_n. This proves part of the theorem.

Transposing the first term of (1) and factoring out a q on the left we prove similarly that q must be a factor of a_0. The student should supply the detailed proof.

Example. Find the rational roots of $f(x) = 4x^4 + 4x^3 + 5x^2 - 10x + 3 = 0$. If p/q is a rational root, reduced to lowest terms, then p must be a factor of 3 and q must be a factor of 4. Hence p can only be ± 1, ± 3 and q can only be ± 1, ± 2, or ± 4. Therefore the only *possible* rational roots are ± 1, ± 3, $\pm 1/2$, $\pm 3/2$, $\pm 1/4$, $\pm 3/4$. Note that instead of possibly having to try infinitely many rational numbers, we may have, at worst, to try 12 possibilities to see if they are roots. In fact, 1/2 is a root r_1 of $f(x) = 0$ since

* We shall not prove this statement here.

$$f(1/2) = 4(1/2)^4 + 4(1/2)^3 + 5(1/2)^2 - 10(1/2) + 3$$
$$= \frac{4}{16} + \frac{4}{8} + \frac{5}{4} - \frac{10}{2} + 3 = 0.$$

Therefore $x - \frac{1}{2}$ is a factor of $f(x)$. By division we obtain $f(x) = (x - \frac{1}{2})(4x^3 + 6x^2 + 8x - 6) = (x - \frac{1}{2})f_1(x)$. Any further roots of $f(x) = 0$ must be roots of the so-called **depressed equation** $f_1(x) = 4x^3 + 6x^2 + 8x - 6 = 0$, obtained by setting the other factor equal to zero. Again $1/2$ is a root r_2 of this equation since $4(1/2)^3 + 6(1/2)^2 + 8(1/2) - 6 = \frac{4}{8} + \frac{6}{4} + 4 - 6 = 0$. By division we find that $f_1(x) = (x - \frac{1}{2})(4x^2 + 8x + 12) = (x - \frac{1}{2})f_2(x)$. Hence, $f(x) = (x - \frac{1}{2})(x - \frac{1}{2})f_2(x)$. The depressed equation $f_2(x) = 4x^2 + 8x + 12 = 0$ is quadratic and can be completely solved, whether or not its roots are rational, by formula. Hence $r_3 = -1 + \sqrt{-2}$ and $r_4 = -1 - \sqrt{-2}$. The number $1/2$ is called a double root, or a root of multiplicity two, of $f(x) = 0$ since $x - \frac{1}{2}$ occurs as a factor of $f(x)$ twice.

Irrational roots may be approximated as closely as we wish by the method of section 96.

EXERCISES

Find all rational roots, and, if the solution yields a quadratic depressed equation, find all the roots:

1. $x^3 - 2x^2 - 5x + 6 = 0$.
2. $x^3 - 4x^2 + x + 6 = 0$.
3. $x^3 - 5x^2 - 8x + 12 = 0$.
4. $x^3 + 3x^2 - 4x - 12 = 0$.
5. $x^4 - x^3 - 19x^2 - 11x + 30 = 0$.
6. $x^4 + 3x^3 - 12x^2 - 13x - 15 = 0$.
7. $9x^4 - 3x^3 + 7x^2 - 3x - 2 = 0$.
8. $3x^3 - 2x^2 + 15x - 10 = 0$.
9. $4x^4 - 7x^2 - 5x - 1 = 0$.
10. $3x^3 - 7x^2 - 3x + 2 = 0$.
11. $6x^4 + 2x^3 + 7x^2 + x + 2 = 0$.
12. $3x^4 + 2x^2 + 5 = 0$.
13. $x^4 - 6x^3 + 10x^2 - 6x + 9 = 0$.
14. $x^4 - 8x^3 + 24x^2 - 32x + 16 = 0$.

171. Permutations. An arrangement of a set of objects in some order in a straight line is called a **permutation** of these objects. More precisely, if we have a set of n objects and we wish

to arrange r of them in some order on a line, each such arrangement is called a **permutation of the n objects taken r at a time.**

Example 1. We have a 3 volume work and a bookshelf with space for only 2 books. The permutations of the 3 volumes of a 3 volume work taken two at a time may be written down as follows (1, 2, 3, denoting volumes 1, 2, 3 respectively): 12, 13, 21, 23, 31, 32.

Example 2. The permutations of four letters a, b, c, d taken 3 at a time are:

abc	*abd*	*acd*	*bcd*
acb	*adb*	*adc*	*bdc*
bac	*bad*	*cad*	*cbd*
bca	*bda*	*cda*	*cdb*
cab	*dab*	*dac*	*dbc*
cba	*dba*	*dca*	*dcb*.

We could arrive at the number of permutations of 3 things taken 2 at a time (example 1) as follows. In the first position (on the bookshelf) we can place any one of the three books. No matter which choice we made for the first position we could put any one of the two remaining books in the second position. By the fundamental principle of page 373, there are $3 \cdot 2$ or 6 different permutations. They are listed in example 1.

Similarly, in example 2, we may place any one of the 4 letters in the first position. No matter which of the 4 choices we make for the first position, we can place any one of the remaining 3 letters in the second position, and then any one of the remaining 2 letters in the third position. By the fundamental principle, there are $4 \cdot 3 \cdot 2 = 24$ permutations. They are listed in example 2.

The number of permutations of n distinct things taken r at a time will be denoted by ${}^nP_r$. Clearly any one of the n things may be placed in the first position, any one of the remaining $(n - 1)$ things may be placed in the second position, any one of the remaining $(n - 2)$ things in the third position, and so on, until we reach the rth position. Hence, by the fundamental principle

$$(1) \qquad {}^nP_r = n(n-1)(n-2) \cdots \qquad \text{(to } r \text{ factors).}$$

Clearly the rth factor will be $n - (r - 1)$ or $n - r + 1$. Hence (1) may be written as

(2) $$^nP_r = n(n - 1)(n - 2) \cdots (n - r + 1).$$

The number of permutations of n distinct objects taken *all* at a time is clearly

(3) $$^nP_n = n(n - 1)(n - 2) \cdots 3 \cdot 2 \cdot 1.$$

We find it convenient to abbreviate the expression $n(n - 1)(n - 2) \cdots 3 \cdot 2 \cdot 1$ by defining the symbol $n\,!$, read "factorial n," to mean the product of all the natural numbers from n down to one; that is,

(4) $$n\,! = n(n - 1)(n - 2) \cdots 3 \cdot 2 \cdot 1.$$

For example $3\,! = 3 \cdot 2 \cdot 1 = 6$, $4\,! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$, $5\,! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$ and so on. Since

$$n(n - 1)(n - 2) \cdots (n - r + 1) = \frac{n(n-1)(n-2)\cdots(n-r+1)\cdot(n-r)(n-r-1)(n-r-2)\cdots 3\cdot 2\cdot 1}{(n-r)(n-r-1)(n-r-2)\cdots 3\cdot 2\cdot 1}$$

$$= \frac{n\,!}{(n - r)\,!}$$

we have the following theorem.

THEOREM. *The number of permutations of n distinct things taken r at a time is given by*

$$^nP_r = \frac{n\,!}{(n - r)\,!} \qquad \textit{if } r < n$$

while

$$^nP_n = n\,!$$

Example 3. The number of arrangements of 7 distinct books on a shelf is $^7P_7 = 7\,! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5040$. The number of arrangements of 7 distinct books taken 4 at a time is

$$^7P_4 = \frac{7\,!}{(7 - 4)\,!} = \frac{7\,!}{3\,!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 7 \cdot 6 \cdot 5 \cdot 4 = 840.$$

EXERCISES

Evaluate each of the following symbols:

1. $6\,!$. **2.** $\frac{5\,!}{3\,!}$. **3.** $\frac{9\,!}{7\,!}$. **4.** $^5P_2$. **5.** $^9P_2$. **6.** $^6P_6$.

7. In how many ways can 5 students be seated in a row of 5 seats?

8. How many different signals can be made from 5 different flags if each signal is to consist of 3 flags hung in a horizontal row?

9. How many different numbers of 2 different digits each can be made from the digits 1, 3, 5, 7, 9?

10. How many different signals can be made from 5 different flags if each signal is to consist of one or more flags hung in a horizontal row?

11. How many different symbols each consisting of 3 letters in succession can be formed from the letters a, b, c, d, e, repetition of a letter in a given symbol being forbidden?

12. How many different symbols each consisting of 3 letters in succession can be formed from the letters a, b, c, d, e, repetitions being permitted?

13. A telephone dial has 10 holes. How many different signals, each consisting of 7 impulses in succession, can be formed, (a) if no impulse is to be repeated in any given signal? (b) if repetitions are permitted?

14. In how many ways may 5 students be seated in a row of 8 seats?

15. In how many ways can the positions of president, vice-president, and secretary be filled in a club of 20 members?

16. If a nickel, dime, quarter, penny, and half-dollar are tossed together, in how many ways may they fall?

17. Three cubical dice are tossed in succession. In how many ways may they fall?

18. In how many ways may a baseball team be arranged in batting orders if a certain four men must occupy the first 4 positions in some order?

19. In how many ways can a party of 5 people be seated in a row of five seats (a) if a certain two insist on sitting next to each other? (b) if the same two refuse to sit next to each other?

20. In how many ways can a set of 4 different mathematics books and 3 different physics books be placed on a shelf if all the books in the same subject must be placed next to each other?

172. Combinations. A set of r objects chosen from a given set of n objects, without regard to order, is called a **combination of n things taken r at a time.**

Example 1. The combinations of 3 distinct volumes taken two at a time are 12, 13, and 23. Compare example 1, section 171. The permutations 12 and 21 give rise to only one combination, since order is to be ignored.

Example 2. The combinations of 4 letters a, b, c, d, taken 3 at a time are abc, abd, acd, bcd. Compare example 2, section 171.

The six permutations in each column of example 2 section 171 give rise to only one combination, since order is to be ignored.

Clearly any combination of n distinct things taken r at a time gives rise to $r\,!$ permutations since the set of r objects can be rearranged among themselves in $r\,!$ ways. Let us denote by ${}^nC_r$ the number of combinations of n distinct things taken r at a time. Then, as we have just seen,

$${}^nP_r = r\,! \cdot {}^nC_r$$

or

$${}^nC_r = \frac{1}{r\,!} \cdot {}^nP_r$$

or

$$(1) \qquad {}^nC_r = \frac{n\,!}{r\,!\,(n-r)\,!} \qquad (r < n).$$

Example 3. In example 1 above, the number of combinations of 3 distinct things taken 2 at a time is ${}^3C_2 = \dfrac{3\,!}{2\,!\,(3-2)\,!} = \dfrac{3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 1} = 3.$ In example 2 above, the number of combinations of 4 distinct things taken 3 at a time is ${}^4C_3 = \dfrac{4\,!}{3\,!\,(4-3)\,!} = \dfrac{4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 1} = 4.$

Remark. Whenever we select a set of r things from a set of n things we automatically select a set of $n - r$ things which are left behind. Therefore ${}^nC_r = {}^nC_{n-r}$.

EXERCISES

Evaluate each of the following symbols:

1. ${}^5C_2$. **2.** ${}^{20}C_{18}$. **3.** ${}^{20}C_2$. **4.** ${}^{17}C_{16} \cdot {}^5C_4 \cdot 3!$

In how many ways can we select:

5. A committee of 3 from a group of 10 people?

6. A jury of 12 men from a panel of 100 eligible men?

7. A set of 3 books from a set of 7 different books?

8. A set of 3 or more books from a set of 7 different books?

9. A set of 3 mathematics books and 2 physics books from a set of 7 mathematics books and 5 physics books, all different?

10. A committee of 4 Democrats and 3 Republicans from a group of 10 Democrats and 8 Republicans?

11. A committee of 5 from a group of 10 (*a*) if a certain two men insist on serving together or not at all? (*b*) if a certain two men refuse to serve together?

12. How many sums of money, each involving three coins, can be formed from a cent, a nickel, a dime, and a quarter?

13. (*a*) How many straight lines are determined by 10 points no 3 of which are in the same straight line? (*b*) How many of these lines pass through any given one of these 10 points?

14. From a group of 12 Democrats and 10 Republicans, how many different committees of 7 can be chosen which contain (*a*) at least 4 Democrats? (*b*) at most 4 Democrats? (*c*) exactly 4 Democrats?

15. In how many different orders can we shelve sets of 5 books, each set consisting of 3 mathematics books and 2 physics books, if the books are to be chosen from a set of 10 mathematics books and 8 physics books, all different?

16. How many triangles are determined by the set of vertices of a regular hexagon?

17. (*a*) How many committees of 5 can be chosen from a group of 15 men? (*b*) How many of these will include a specified man A? (*c*) From how many will A be excluded?

18. Prove that ${}^{n-1}C_r + {}^{n-1}C_{r-1} = {}^nC_r$.

173. The binomial theorem. By tedious multiplication we find that

$$
\begin{aligned}
(x+h)^1 &= x + h \\
(x+h)^2 &= x^2 + 2\,xh + h^2 \\
(x+h)^3 &= x^3 + 3\,x^2h + 3\,xh^2 + h^3 \\
(x+h)^4 &= x^4 + 4\,x^3h + 6\,x^2h^2 + 4\,xh^3 + h^4 \\
(x+h)^5 &= x^5 + 5\,x^4h + 10\,x^3h^2 + 10\,x^2h^3 + 5\,xh^4 + h^5,
\end{aligned}
$$

and so on. It is clear that the expansion of $(x + h)^n$, where n is any natural number, begins with x^n and ends with h^n. In the intermediate terms the exponent of x decreases by one with each successive term and the exponent of h increases by one, so that the sum of the two exponents is exactly n in each term. The coefficients may be obtained from the following scheme, known as **Pascal's triangle**, which we shall not justify here,

$$
\begin{array}{ccccccccccc}
 & & & & 1 & & 1 & & & & \\
 & & & 1 & & 2 & & 1 & & & \\
 & & 1 & & 3 & & 3 & & 1 & & \\
 & 1 & & 4 & & 6 & & 4 & & 1 & \\
1 & & 5 & & 10 & & 10 & & 5 & & 1 \\
\end{array}
$$

. . . .

in which each number, except the 1's on the outside, is the sum of the two nearest numbers in the line above. Hence

$$(x + h)^6 = x^6 + 6\,x^5h + 15\,x^4h^2 + 20\,x^3h^3 + 15\,x^2h^4 + 6\,xh^5 + h^6.$$

This scheme is, however, not very practical for high exponents n, for to write even the first few terms of $(x + h)^{175}$, say, we should have to work out Pascal's triangle down to the 175th line. A better scheme is given by the following theorem.

Theorem. *If n is a natural number*

$$(x + h)^n = x^n + {}^nC_1x^{n-1}h + {}^nC_2x^{n-2}h^2 + {}^nC_3x^{n-3}h^3 + \cdots + {}^nC_rx^{n-r}h^r + \cdots + h^n.$$

This result appears natural if you recall that $(x + h)^n$ means $(x + h)(x + h) \cdots (x + h)$ with n factors. Each term in the result is a sum of terms each of which is a product of one letter from each parenthesis. For example, the term involving h^r will be a sum of terms of the form $x^{n-r}h^r$ each of which is a product of $n - r$ x's (one from each of $n - r$ parentheses) and r h's (one from each of r parentheses). How many such terms $x^{n-r}h^r$ will there be? Clearly, as many as there are choices of r of the n parentheses from which to take an h. Hence there are ${}^nC_r$ such terms and the total coefficient of $x^{n-r}h^r$ in the final result is ${}^nC_r$.

Example.
$$\begin{aligned}(x + h)^6 &= x^6 + {}^6C_1x^5h + {}^6C_2x^4h^2 + {}^6C_3x^3h^3 + {}^6C_4x^2h^4 + {}^6C_5xh^5 + h^6\\ &= x^6 + 6\,x^5h + 15\,x^4h^2 + 20\,x^3h^3 + 15\,x^2h^4 + 6\,xh^5 + h^6.\end{aligned}$$

This theorem, known as the **binomial theorem for positive integral exponents,** can be proved by mathematical induction. This will not be done here.

The coefficient of the term involving h^r is ${}^nC_r$; this is the $(r + 1)$th term in the expansion. There will be $n + 1$ terms, in all, in the expansion of $(x + h)^n$.

EXERCISES

Write the expansion of each of the following:

1. $(x + h)^7$. **2.** $(a + b)^8$. **3.** $(p + q)^9$.

4. $(2\,a + 3\,b)^4$. **5.** $(2\,a^3 + 3\,b^2)^5$. **6.** $\left(\frac{a}{2} + \frac{4\,b^2}{a}\right)^6$.

7. $(a - b)^3$. (Hint: write $a - b = a +[- b]$.)

8. $(a - b)^4$. **9.** $(a - b)^5$. **10.** $\left(2a^3 - \dfrac{b^2}{2a}\right)^6$.

11. $\left(1 + \dfrac{1}{2}\right)^2$. **12.** $\left(1 + \dfrac{1}{3}\right)^3$. **13.** $\left(1 + \dfrac{1}{4}\right)^4$.

14. $\left(1 + \dfrac{1}{5}\right)^5$.

Write, in simplified form, only the specified terms:

15. The term involving h^3 in the expansion of $(x + h)^{17}$.

16. The term involving b^5 in the expansion of $(a + b)^{24}$.

17. The sixth term of $(x + h)^{15}$.

18. The 18th term of $(x + h)^{19}$.

19. The term involving b^6 in the expansion of $\left(\dfrac{a^2}{2} + 2b^2\right)^{10}$.

20. The term involving q^{10} in the expansion of $\left(\dfrac{p^2}{3} + 6q^2\right)^{12}$.

21. The middle term of $(a - b)^8$.

22. The middle terms of $(2a - 3b)^{11}$.

23. The term involving b^3 in $\left(\dfrac{b}{a} + \dfrac{1}{b}\right)^9$.

24. Explain the connection between exercise 18 of section 172 and Pascal's triangle.

25. Compute, using only enough terms to get the result accurate to 3 decimal places:

(*a*) $1.01^9 = (1 + .01)^9$.
(*b*) $.99^6 = (1 - .01)^6$.
(*c*) 1.02^{11}.
(*d*) $.98^7$.

174. Further results on probability. We discuss briefly some elementary theorems on a priori probability.

THEOREM 1. *If the probability that an event will happen is p then the probability that it will not happen is 1 − p.*

For if the event can turn out successfully in s ways out of a total number t of possible (equally likely) ways, then it will fail in the remaining $t - s$ ways. Hence the probability of failure is $\dfrac{t-s}{t} = \dfrac{t}{t} - \dfrac{s}{t} = 1 - p$ where $p = \dfrac{s}{t}$ is the probability of success.

Example 1. The probability of throwing a 3 with a single die is $1/6$ while the probability of failing to throw a 3 is $1 - 1/6 = 5/6$.

Two or more events are called **independent** if the occurrence or non-occurrence of any one of them is not affected by the occurrence or non-occurrence of the rest.

THEOREM 2. *If p_1 is the probability that a first event will occur, p_2 is the probability that a second event will occur, p_3 is the probability that a third event will occur, and so on, and if all these events are independent, then the probability that* ***all*** *of them will occur is $p_1p_2p_3 \cdots$.*

For simplicity, suppose there are two events with probabilities $p_1 = s_1/t_1$ and $p_2 = s_2/t_2$ respectively. By the fundamental principle of page 373, any of the t_1 possibilities for the first event may be combined with any one of the t_2 possibilities for the second event, yielding a total number t_1t_2 of possibilities for both events. By the same reasoning the total number of successful cases for both events together is s_1s_2. Hence the probability of both events occurring successfully is $\frac{s_1s_2}{t_1t_2} = \frac{s_1}{t_1} \cdot \frac{s_2}{t_2} = p_1p_2$. The proof for more than two events is similar.

Example 2. The probability that an ace be drawn from a pack of cards is 4/52. Suppose that after a card is drawn, it is replaced in the pack before drawing again. Then the probability of drawing two aces in succession is $\frac{4}{52} \cdot \frac{4}{52} = \frac{1}{169}$ since the two drawings are independent of each other.

THEOREM 3. *If the probability of a first event is p_1 and if the probability of a second event, after the first event has happened, is p_2, then the probability that both events will occur in succession is p_1p_2. A similar theorem can be stated for more than two events.*

This is proved in the same way as Theorem 2.

Example 3. Suppose that when a card is drawn from a pack it is not to be replaced. Then the probability of drawing two aces in succession is $\frac{4}{52} \cdot \frac{3}{51} = \frac{1}{221}$ since, after the first ace has been drawn and not replaced, the probability of drawing a second ace is 3/51.

Two or more events are said to be **mutually exclusive** if not more than one of them can occur.

THEOREM 4. *If the probability that one event will occur is p_1 and the probability that a second event will occur is p_2, and if the two events are mutually exclusive, then the probability that either the first or the second event will occur is $p_1 + p_2$. A similar theorem can be stated for more than two events.*

Let the total number of possibilities be t, of which s_1 cases are favorable for the first event and s_2 cases are favorable for the second event. The s_1 cases and the s_2 cases are all different since the two events are mutually exclusive. The probability of the first event is $p_1 = s_1/t$ and the probability of the second event is $p_2 = s_2/t$. One or the other of the two events can happen in $s_1 + s_2$ cases. Therefore the probability that either one or the other of the two events will occur is $\dfrac{s_1 + s_2}{t} = \dfrac{s_1}{t} + \dfrac{s_2}{t} = p_1 + p_2$.

Example 4. The probability that either an ace or a king will be drawn in a single draw from a pack of cards is $\dfrac{4}{52} + \dfrac{4}{52} = \dfrac{2}{13}$.

THEOREM 5. *If p is the probability that an event will occur in a single trial, then the probability that it will occur exactly r times out of n trials is ${}^nC_r q^{n-r} p^r$ where $q = 1 - p$ is the probability that the event will fail to occur in a single trial.*

The probability that the event will occur in any particular set of r trials and fail in the remaining $n - r$ trials is exactly $q^{n-r}p^r$ by theorem 2. But the particular set of r trials may be selected out of n trials in ${}^nC_r$ ways which are mutually exclusive. Hence, by theorem 4, we must add ${}^nC_r$ terms each of which has the value $q^{n-r}p^r$. Therefore the probability desired is ${}^nC_r q^{n-r} p^r$.

Example 5. Find the probability of obtaining exactly two heads in tossing 5 coins. The probability of obtaining a head with any particular coin is $p = \dfrac{1}{2}$ and the probability of failing to obtain a head with any particular coin is $q = 1 - \dfrac{1}{2} = \dfrac{1}{2}$. Hence the probability in question is ${}^5C_2\left(\dfrac{1}{2}\right)^3\left(\dfrac{1}{2}\right)^2 = \dfrac{5}{16}$.

EXERCISES

1. What is the probability of throwing an ace twice in succession with a single die?

2. What is the probability of throwing a total of 18 with 3 dice?

3. What is the probability of drawing an ace from a pack of 52 cards four times in succession if: (*a*) each card is replaced in the pack after it is drawn? (*b*) no card is replaced after it is drawn?

4. What is the probability that either an ace or a six will turn up in a single throw of a single die?

5. What is the probability that either an ace, king, or queen will be drawn from a pack of cards in a single draw?

6. A bag contains 6 white and 5 black balls. After drawing any ball it is to be replaced in the bag. What is the probability of drawing two white balls in succession? Three black balls in succession?

7. The same as exercise 6 except that no ball is to be replaced after drawing.

8. What is the probability that an ace will turn up exactly once in 3 successive throws of a single die?

9. What is the probability that an ace will turn up exactly twice in 5 successive throws of a single die?

10. What is the probability that an ace will turn up at least once in four successive throws of a single die? (Hint: consider the probability of failure.)

11. If 10 coins are tossed in succession, what is the probability that (*a*) exactly 3 will be heads? (*b*) at least 3 will be heads?

12. If 10 coins are tossed, what is the probability that (*a*) exactly 8 will be heads? (*b*) at least 8 will be heads? (*c*) at most 8 will be heads?

13. If 5 dice are tossed what is the probability that (*a*) exactly 3 of them will turn up an ace? (*b*) at least 3 of them will turn up an ace? (*c*) at most 3 of them will turn up an ace?

14. What is the probability of drawing a hand consisting of the ace, king, queen, jack and ten of diamonds in 5 successive draws from a pack of 52 cards?

15. What is the probability of throwing a total of seven with a pair of dice (*a*) 5 times in succession? (*b*) ten times in succession? (*c*) is the probability in (*b*) half as great as that in (*a*)?

Table I

Common Logarithms

N	0	1	2	3	4	5	6	7	8	9
10	0000	0043	0086	0128	0170	0212	0253	0294	0334	0374
11	0414	0453	0492	0531	0569	0607	0645	0682	0719	0755
12	0792	0828	0864	0899	0934	0969	1004	1038	1072	1106
13	1139	1173	1206	1239	1271	1303	1335	1367	1399	1430
14	1461	1492	1523	1553	1584	1614	1644	1673	1703	1732
15	1761	1790	1818	1847	1875	1903	1931	1959	1987	2014
16	2041	2068	2095	2122	2148	2175	2201	2227	2253	2279
17	2304	2330	2355	2380	2405	2430	2455	2480	2504	2529
18	2553	2577	2601	2625	2648	2672	2695	2718	2742	2765
19	2788	2810	2833	2856	2878	2900	2923	2945	2967	2989
20	3010	3032	3054	3075	3096	3118	3139	3160	3181	3201
21	3222	3243	3263	3284	3304	3324	3345	3365	3385	3404
22	3424	3444	3464	3483	3502	3522	3541	3560	3579	3598
23	3617	3636	3655	3674	3692	3711	3729	3747	3766	3784
24	3802	3820	3838	3856	3874	3892	3909	3927	3945	3962
25	3979	3997	4014	4031	4048	4065	4082	4099	4116	4133
26	4150	4166	4183	4200	4216	4232	4249	4265	4281	4298
27	4314	4330	4346	4362	4378	4393	4409	4425	4440	4456
28	4472	4487	4502	4518	4533	4548	4564	4579	4594	4609
29	4624	4639	4654	4669	4683	4698	4713	4728	4742	4757
30	4771	4786	4800	4814	4829	4843	4857	4871	4886	4900
31	4914	4928	4942	4955	4969	4983	4997	5011	5024	5038
32	5051	5065	5079	5092	5105	5119	5132	5145	5159	5172
33	5185	5198	5211	5224	5237	5250	5263	5276	5289	5302
34	5315	5328	5340	5353	5366	5378	5391	5403	5416	5428
35	5441	5453	5465	5478	5490	5502	5514	5527	5539	5551
36	5563	5575	5587	5599	5611	5623	5635	5647	5658	5670
37	5682	5694	5705	5717	5729	5740	5752	5763	5775	5786
38	5798	5809	5821	5832	5843	5855	5866	5877	5888	5899
39	5911	5922	5933	5944	5955	5966	5977	5988	5999	6010
40	6021	6031	6042	6053	6064	6075	6085	6096	6107	6117
41	6128	6138	6149	6160	6170	6180	6191	6201	6212	6222
42	6232	6243	6253	6263	6274	6284	6294	6304	6314	6325
43	6335	6345	6355	6365	6375	6385	6395	6405	6415	6425
44	6435	6444	6454	6464	6474	6484	6493	6503	6513	6522
45	6532	6542	6551	6561	6571	6580	6590	6599	6609	6618
46	6628	6637	6646	6656	6665	6675	6684	6693	6702	6712
47	6721	6730	6739	6749	6758	6767	6776	6785	6794	6803
48	6812	6821	6830	6839	6848	6857	6866	6875	6884	6893
49	6902	6911	6920	6928	6937	6946	6955	6964	6972	6981
50	6990	6998	7007	7016	7024	7033	7042	7050	7059	7067
51	7076	7084	7093	7101	7110	7118	7126	7135	7143	7152
52	7160	7168	7177	7185	7193	7202	7210	7218	7226	7235
53	7243	7251	7259	7267	7275	7284	7292	7300	7308	7316
54	7324	7332	7340	7348	7356	7364	7372	7380	7388	7396
N	0	1	2	3	4	5	6	7	8	9

Table I (Cont.)

N	0	1	2	3	4	5	6	7	8	9
55	7404	7412	7419	7427	7435	7443	7451	7459	7466	7474
56	7482	7490	7497	7505	7513	7520	7528	7536	7543	7551
57	7559	7566	7574	7582	7589	7597	7604	7612	7619	7627
58	7634	7642	7649	7657	7664	7672	7679	7686	7694	7701
59	7709	7716	7723	7731	7738	7745	7752	7760	7767	7774
60	7782	7789	7796	7803	7810	7818	7825	7832	7839	7846
61	7853	7860	7868	7875	7882	7889	7896	7903	7910	7917
62	7924	7931	7938	7945	7952	7959	7966	7973	7980	7987
63	7993	8000	8007	8014	8021	8028	8035	8041	8048	8055
64	8062	8069	8075	8082	8089	8096	8102	8109	8116	8122
65	8129	8136	8142	8149	8156	8162	8169	8176	8182	8189
66	8195	8202	8209	8215	8222	8228	8235	8241	8248	8254
67	8261	8267	8274	8280	8287	8293	8299	8306	8312	8319
68	8325	8331	8338	8344	8351	8357	8363	8370	8376	8382
69	8388	8395	8401	8407	8414	8420	8426	8432	8439	8445
70	8451	8457	8463	8470	8476	8482	8488	8494	8500	8506
71	8513	8519	8525	8531	8537	8543	8549	8555	8561	8567
72	8573	8579	8585	8591	8597	8603	8609	8615	8621	8627
73	8633	8639	8645	8651	8657	8663	8669	8675	8681	8686
74	8692	8698	8704	8710	8716	8722	8727	8833	8839	8845
75	8751	8756	8762	8768	8774	8779	8785	8791	8797	8802
76	8808	8814	8820	8825	8831	8837	8842	8848	8854	8859
77	8865	8871	8876	8882	8887	8893	8899	8904	8910	8915
78	8921	8927	8932	8938	8943	8949	8954	8960	8965	8971
79	8976	8982	8987	8993	8998	9004	9009	9015	9020	9025
80	9031	9036	9042	9047	9053	9058	9063	9069	9074	9079
81	9085	9090	9096	9101	9106	9112	9117	9122	9128	9133
82	9138	9143	9149	9154	9159	9165	9170	9175	9180	9186
83	9191	9196	9201	9206	9212	9217	9222	9227	9232	9238
84	9243	9248	9253	9258	9263	9269	9274	9279	9284	9289
85	9294	9299	9304	9309	9315	9320	9325	9330	9335	9340
86	9345	9350	9355	9360	9365	9370	9375	9380	9385	9390
87	9395	9400	9405	9410	9415	9420	9425	9430	9435	9440
88	9445	9450	9455	9460	9465	9469	9474	9479	9484	9489
89	9494	9499	9504	9509	9513	9518	9523	9528	9533	9538
90	9542	9547	9552	9557	9562	9566	9571	9576	9581	9586
91	9590	9595	9600	9605	9609	9614	9619	9624	9628	9633
92	9638	9643	9647	9652	9657	9661	9666	9671	9675	9680
93	9685	9689	9694	9699	9703	9708	9713	9717	9722	9727
94	9731	9736	9741	9745	9750	9754	9759	9763	9768	9773
95	9777	9782	9786	9791	9795	9800	9805	9809	9814	9818
96	9823	9827	9832	9836	9841	9845	9850	9854	9859	9863
97	9868	9872	9877	9881	9886	9890	9894	9899	9903	9908
98	9912	9917	9921	9926	9930	9934	9939	9943	9948	9952
99	9956	9961	9965	9969	9974	9978	9983	9987	9991	9996
N	0	1	2	3	4	5	6	7	8	9

Table II

Trigonometric Functions

	sin	cos	tan	cot	sec	csc	
0°	.0000	1.0000	.0000	....	1.000	....	90°
1°	.0175	.9998	.0175	57.29	1.000	57.30	89°
2°	.0349	.9994	.0349	28.64	1.001	28.65	88°
3°	.0523	.9986	.0524	19.08	1.001	19.11	87°
4°	.0698	.9976	.0699	14.30	1.002	14.34	86°
5°	.0872	.9962	.0875	11.43	1.004	11.47	85°
6°	.1045	.9945	.1051	9.514	1.006	9.567	84°
7°	.1219	.9925	.1228	8.144	1.008	8.206	83°
8°	.1392	.9903	.1405	7.115	1.010	7.185	82°
9°	.1564	.9877	.1584	6.314	1.012	6.392	81°
10°	.1736	.9848	.1763	5.671	1.015	5.759	80°
11°	.1908	.9816	.1944	5.145	1.019	5.241	79°
12°	.2079	.9781	.2126	4.705	1.022	4.810	78°
13°	.2250	.9744	.2309	4.331	1.026	4.445	77°
14°	.2419	.9703	.2493	4.011	1.031	4.134	76°
15°	.2588	.9659	.2679	3.732	1.035	3.864	75°
16°	.2756	.9613	.2867	3.487	1.040	3.628	74°
17°	.2924	.9563	.3057	3.271	1.046	3.420	73°
18°	.3090	.9511	.3249	3.078	1.051	3.236	72°
19°	.3256	.9455	.3443	2.904	1.058	3.072	71°
20°	.3420	.9397	.3640	2.747	1.064	2.924	70°
21°	.3584	.9336	.3839	2.605	1.071	2.790	69°
22°	.3746	.9272	.4040	2.475	1.079	2.669	68°
23°	.3907	.9205	.4245	2.356	1.086	2.559	67°
24°	.4067	.9135	.4452	2.246	1.095	2.459	66°
25°	.4226	.9063	.4663	2.145	1.103	2.366	65°
26°	.4384	.8988	.4877	2.050	1.113	2.281	64°
27°	.4540	.8910	.5095	1.963	1.122	2.203	63°
28°	.4695	.8829	.5317	1.881	1.133	2.130	62°
29°	.4848	.8746	.5543	1.804	1.143	2.063	61°
30°	.5000	.8660	.5774	1.732	1.155	2.000	60°
31°	.5150	.8572	.6009	1.664	1.167	1.942	59°
32°	.5299	.8480	.6249	1.600	1.179	1.887	58°
33°	.5446	.8387	.6494	1.540	1.192	1.836	57°
34°	.5592	.8290	.6745	1.483	1.206	1.788	56°
35°	.5736	.8192	.7002	1.428	1.221	1.743	55°
36°	.5878	.8090	.7265	1.376	1.236	1.701	54°
37°	.6018	.7986	.7536	1.327	1.252	1.662	53°
38°	.6157	.7880	.7813	1.280	1.269	1.624	52°
39°	.6293	.7771	.8098	1.235	1.287	1.589	51°
40°	.6428	.7660	.8391	1.192	1.305	1.556	50°
41°	.6561	.7547	.8693	1.150	1.325	1.524	49°
42°	.6691	.7431	.9004	1.111	1.346	1.494	48°
43°	.6820	.7314	.9325	1.072	1.367	1.466	47°
44°	.6947	.7193	.9657	1.036	1.390	1.440	46°
45°	.7071	.7071	1.000	1.000	1.414	1.414	45°
	cos	sin	cot	tan	csc	sec	

ANSWERS

Mostly to Odd-Numbered Exercises

Section 3

1. $3\frac{3}{4}$ days. **3.** Less. **5.** The second. In successive six month periods the first receives 900, 900, 1000, 1000, 1100, 1100, . . . dollars, while the second receives 900, 950, 1000, 1050, 1100, 1150, . . . dollars.

Section 4

1. Not valid. **3.** Not valid. **5.** Not valid. **7.** Valid. **9.** Valid. **11.** (*a*) Not valid. (*b*) Not valid. (*c*) Not valid. **15.** Yes. **17.** No.

Section 5

1. Not valid. **3.** Not valid. **5.** Not valid. **7.** Not valid. **9.** Not valid.

Section 12

1. (*a*) A_2. (*b*) A_3. (*c*) A_2. (*d*) A_3. (*e*) A_1.

Section 13

1. (*a*) M_2. (*b*) M_3. (*c*) D. (*d*) D and E_2. (*e*) M_2. (*f*) M_3. (*g*) D. (*h*) D and E_2. **3.** (*a*) 23. (*b*) 35. (*c*) 17. (*d*) 27. (*e*) 29. (*f*) 41. (*g*) 77. (*h*) 47.

Section 15

3. (*b*), (*c*), (*d*), (*e*).

Section 16

1. (*a*) 10/21. (*b*) 14/15. (*c*) 1. (*d*) 3/2. (*e*) 1. (*f*) 1/2. **3.** ad/bd.

Section 17

1. (*a*) 13/6. (*b*) 5/6. (*e*) 2. (*f*) 2. **5.** Less.

Section 18

1. (*a*) 13. (*b*) 1. (*c*) $1 + 2x$. (*d*) 7/12. (*e*) 5. (*f*) $\dfrac{ad + bc}{ad - bc}$.

Section 19

1. None. **3.** Symmetric. **5.** Transitive. **7.** None. **9.** If "A is west of B" means that A can be reached by proceeding west from B, then the relation is reflexive, symmetric, and transitive. If "A is west of B" means that the shortest route from B to A is to the west, then it is none. **11.** Reflexive, transitive. **13.** Symmetric. **15.** Reflexive, symmetric, transitive. **17.** Reflexive, symmetric.

Section 21

1. 3.	**3.** -3.	**5.** -3.	**7.** -10.
9. 10.	**11.** 4.	**15.** -1.	**17.** 1/6.
19. $-7/6$.	**21.** 3/4.	**23.** $-1/3$.	**25.** (*a*) $-a + b$. (*b*) $a - b$.

Section 23

1. (*a*) 3. (*b*) 2. (*c*) 7. (*d*) 2 (*e*) 2. (*f*) 3. **3.** No. **5.** $(a - b)^2 = a^2 - 2ab + b^2$. **7.** No. **9.** 1/2. **11.** 1/2. **13.** x^2. **15.** 36. **17.** 23. **19.** 41 **21.** 441. **23.** $6x$. **25.** $3x$.

Section 25

1. (*a*) between 1.732 and 1.733. (*b*) 1.73.
3. (*a*) between 1.259 and 1.260. (*b*) 1.26.
5. (*a*) between 1.710 and 1.711. (*b*) 1.71.

Section 26

1. (*a*) complex, imaginary.
(*b*) complex, real, rational, integral, positive, natural.
(*c*) complex, imaginary, pure imaginary.
(*d*) complex, real, irrational, positive.
(*e*) complex, real, irrational, positive.
(*f*) complex, real, rational, negative.
(*g*) complex, real, irrational, positive.
(*h*) complex, real, rational, integral, positive, natural.

Section 29

1. (*a*) Algebraic, irrational. (*b*) 4.
3. (*a*) Algebraic, rational, polynomial. (*b*) $\frac{3}{4} + \sqrt{3}$.
5. (*a*) Algebraic, rational, polynomial. (*b*) 3.
7. (*a*) Algebraic, rational. (*b*) 3/10.
9. (*a*) Algebraic, rational, polynomial. (*b*) 59.
11. Rational.
13. No.

Section 30

1. $5x^2 + 5x + 6$. **3.** $7x^3 + 2x^2 + 5x + 7$. **5.** $x^2 + 5x + 6$.
7. $4a^2x^2 + 4abx + b^2$. **9.** $x^2 - 2xy + y^2$. **11.** $7x + 2xy - 3y^2$. **13.** $x^3 + y^3$.

Section 31

1. $2xy(2x + 3y - 5x^2y^2)$. **3.** $(x - 5)(x - 2)$. **5.** $(x + 3)(x - 3)$.
7. $(2x + 1)(x + 2)$. **9.** $(y + k)^2$. **11.** $(2ax + b)^2$.

Section 32

9. 3. **11.** 2. **13.** Conditional. **15.** Conditional. **17.** Identity.

Section 33

1. $-13/4$. **3.** 12. **5.** 1. **7.** No root.

Section 34

1. 5,2. **3.** $-4/3$,5. **5.** ± 3. **7.** -7,1. **9.** $\pm 5/2$. **11.** 0.

Section 35

1. 7. **3.** 4. **5.** 4,20. **7.** No root. **9.** -2. **11.** 9/8. **13.** No root.

Section 36

1. 3,2; real, rational. **3.** $\frac{5 \pm \sqrt{21}}{2}$; real, irrational.
5. $\pm\sqrt{-3}$; imaginary. **7.** 5, $-4/3$; real, rational.
9. $-3 \pm \sqrt{17}$; real, irrational. **11.** $\frac{-7 \pm 3\sqrt{5}}{2}$; real, irrational.

Section 39

1. $x = 2$, $y = 3$. **3.** $x = -2$, $y = -3$. **5.** $x = 8/5$, $y = -7/10$.
7. $x = 9/13$, $y = 7/13$. **9.** $x = 14/29$, $y = -23/29$.

Section 40

1. $x = 5, y = 0; x = 3, y = 4$.
3. $x = 3, y = 2; x = 1/5, y = -18/5$.
5. $x = -13/3, y = -4; x = -11/9, y = 2/3$.
7. $x = 3, y = 4; x = 4, y = 3$.
9. $x = 3, y = 4; x = -3, y = -4$.
11. $x = 3, y = 2; x = 21/4, y = -19/4$.

Section 41

1. 8 yds., 6 yds. **3.** $3\frac{3}{4}$. **5.** 2. **7.** \$1000 at 4%, \$4000 at 6%. **9.** 8 days. **11.** 40 ft., 10 ft. **13.** A: 10 days; B: 15 days. **15.** 40 ft., 10 ft. **17.** 2 hrs., 100 mi. **19.** $5\frac{5}{8}$ gallons.

Section 44

5. 4. **7.** 2. **9.** 4. **11.** 2. **19.** 2, 5. **21.** 1, 3, 5. **23.** 4.

Section 47

1. (*a*) 24. (*b*) 44. (*c*) 26. (*d*) 30. (*e*) 33. (*f*) 40. (*g*) 120. (*h*) 220. (*i*) 11000.
3. (*a*) 32. (*b*) 112. (*c*) 35. (*d*) 40. (*e*) 44. (*f*) 52. (*g*) 200. (*h*) 1012. (*i*) 100000.
5. (*a*) 8. (*b*) 13. (*c*) 8. (*d*) 10. (*e*) 11. (*f*) 12. (*g*) 20. (*h*) 22. (*i*) 1000.
7. (*a*) 62. (*b*) 146. (*c*) 26. (*d*) 222.
9. (*a*) $2 + 10 = 12$. (*b*) $10 + 101 = 111$. (*c*) $2 + 12 = 21$. (*d*) $2 + 5 = 10$. (*e*) $2 + 5 = 7$.
15. (*a*) 54. (*b*) 41. (*c*) $t0$. (*d*) te. (*e*) $11e$.

Section 49

1. a^8. **3.** a^3/b^4. **5.** 72. **7.** -72.
9. 18. **11.** 36. **13.** 3^5. **15.** 3^6.

Section 50

1. $1/9$. **3.** 8. **5.** 10.
7. 1. **9.** $9/100$. **11.** 25,600,000.
13. $3/x^3$. **15.** 3. **17.** $8b^2$.
19. $2x^6y^4$. **21.** $x^3 + 3x + 5$. **23.** $1/x^3y^5$.

Section 51

1. 2×10^5. **3.** 5.67×10^{11}. **5.** 5.6×10^{-2}. **7.** 8.93×10^{-8}. **9.** 4.678×10^1. **11.** .1. **13.** 67,800,000. **15.** 6.78. **17.** 468,000. **19.** 90. **21.** 5×10^{14}. **23.** 220. **25.** 1.316385×10^{25} lbs.

Section 52

1. 4. **3.** 2. **5.** $1/4$. **7.** $1/4$. **9.** $1/64$. **11.** $1/243$. **13.** $1/3$. **15.** $\sqrt[28]{x^{41}}$. **17.** $\sqrt[15]{x}$.

Section 53

1. $\log_5 25 = 2$. **3.** $\log_{10} 100 = 2$. **5.** $\log_8 2 = 1/3$. **7.** $\log_5(1/25) = -2$. **9.** $\log_7 1 = 0$. **11.** $2^4 = 16$. **13.** $10^{-2} = .01$. **15.** $10^0 = 1$. **17.** $5^3 = 125$. **19.** $10^{-1} = .1$. **21.** 5. **23.** $2/3$. **25.** 1. **27.** 0. **29.** $3/2$. **31.** 4. **33.** 5. **35.** .7781. **37.** .1761. **39.** 1.0791. **41.** $-.6990$.

Section 54

1. 2.0934. **3.** 8.0934–10. **5.** 1.5977. **7.** 8.6385–10. **9.** 6.5391–10. **11.** 2.5563. **13.** 7.6990–10. **15.** 33.2. **17.** .332. **19.** 49.8. **21.** .545. **23.** 9370. **25.** $10^{2.9355} = 862$; $\sqrt[10000]{10^{29355}} = 862$.

Section 55

1. 1.72. **3.** 2.59. **5.** .783. **7.** 40.7. **9.** 1280. **11.** .205.

Section 56

1. \$184. **3.** \$5,230,000. **5.** 197,000,000 sq. mi. **7.** 18 yrs. **9.** 2 sec. **11.** (*a*) \$328. (*b*) \$107.

Section 71

1. (*a*) 2. (*b*) 6. **3.** (*a*) 4. (*b*) -5. **5.** (*a*) 10. (*b*) 2. **7.** (*a*) 8. (*b*) 4. **9.** (*a*) 3. (*b*) 7/2. **11.** 7.

Section 72

5. 0. **7.** Equal. **9.** $(3,-3)$, $(-3,-3)$, $(-3,3)$.

Section 74

1. 13. **3.** 4. **5.** 10. **7.** 4. **13.** (*b*) 5. (*c*) No. (*d*) No. (*e*) Yes. **15.** $(0,-3)$.

Section 75

1. (3,4). **3.** (*a*) $7, \frac{1}{2}\sqrt{85}$, $\frac{1}{2}\sqrt{85}$; (*c*) $\sqrt{26}$, $\frac{1}{2}\sqrt{221}$, $\frac{1}{2}\sqrt{65}$. **5.** $(-5,-1)$. **7.** $\left(\frac{a+c}{2}, \frac{b+d}{2}\right)$.

Section 76

3. (*a*) $-1/5$, $-7/9$, 5. (*c*) $-5/12$, 0, does not exist.

Section 78

1. (*a*) Yes. (*b*) No. (*c*) Yes. **15.** $y = -3$. **17.** $x = y$. **19.** $y = 2x$.

Section 79

1. $2x - y = 9$. **3.** $y = 2$. **5.** $y = x + 1$. **7.** $x = 2$. **9.** $y = 3x + 4$. **11.** $y = -x$. **13.** -1. **15.** $m = 3/2$, $p = 1/2$.

Section 80

1. 2/3. **3.** No slope. **5.** $-1/2$. **7.** $y = 0$; $x = 0$. **9.** $2x + y = 8$.
11. (*a*) $y = 4x + 8$; $3y = x + 2$; $3x + 2y = 16$.
(*b*) $3y = 2x + 4$; $x + 4y = 12$; $3x + y = 8$.
(*c*) $4y = 5x + 10$; $7x + y = 8$; $2x + 5y = 18$.
(*d*) $x + 4y = 15$; $y + 3x = 4$; $3y = 2x + 11$.
13. $x + y = 5$; no; yes. **15.** $x^2 + y^2 = 25$; yes; no.

Section 81

1. $(x + 1)^2 + (y - 3)^2 = 16$. **3.** $x^2 + y^2 = 4$. **5.** $(x + 2)^2 + y^2 = 4$. **7.** $(x + 2)^2 + (y + 3)^2 = 169$. **9.** $(2,-5)$; 4. **11.** (1,0); 3. **13.** $x^2 + y^2 = 25$; yes; yes. **15.** $x^2 + y^2 = 25$; yes; yes.

Section 82

7. $y^2 = 8x$. **9.** $9x^2 - 16y^2 = 144$.

Section 83

1. $x = y$. **3.** $4y - 4x = 15$.

Section 84

1. $(-1,-2)$. **3.** $(23/17\ \ -1/17)$. **5.** (1,2). **7.** (3,4), $(-3,-4)$. **9.** (4,5), (5,4). **11.** (1,1), $(9/16, -3/4)$. **12.** No intersections. **13.** (0,1), (3,4). **15.** $(-3,2)$, $(-1,-2)$. **17.** $(-3,-2)$, $(-1,2)$. **21.** (*a*) (3,0), (4,2), (1,2). (*b*) $2y = x$; $2x + 5y = 12$; $4x + y = 12$. (*c*) (8/3,4/3). **23.** (*a*) $x = 3$, $x = y + 2$, $x + 2y = 5$.

Section 90

1. All real numbers. **3.** All non-negative real numbers. **5.** All real numbers between -3 and $+3$ inclusive.

Section 91

1. (*a*) 4. (*b*) 3. (*c*) 6. (*d*) 18. (*e*) 3.
3. (*a*) $-7/4$. (*b*) $-1/2$. (*c*) 17/10. (*d*) $-4/3$. (*e*) 2,5.
5. (*a*) 19. (*b*) $3a^2 + 2a + 3$. (*c*) $3a^2 + 6ah + 3h^2 + 2a + 2h + 3$. (*d*) $6ah + 3h^2 + 2h$. (*e*) $6a + 3h + 2$.
7. $3x^2 + 3xh + h^2$.
9. $-1/(x^2 + xh)$.
11. $-(2x + h)/x^2(x + h)^2$.

Section 92

7. (*a*) $A = 20x - x^2$. **9.** (*a*) $A = 2x^2 + 256x^{-1}$. **11.** (*a*) $V = 84x^2 - 4x^3$.

Section 93

1. (*a*) 80 ft. per sec. (*b*) 112 ft. per sec. (*c*) 48 ft. per sec. (*d*) 64 ft. per sec.

Section 94

1. (*a*) 9.5. (*b*) 9.3. (*c*) 9.5. (*d*) 3.5. (*e*) 2. **3.** (*a*) 1.1804. (*b*) 1.1796. (*c*) 1.1797. (*d*) 1.1811. **5.** (*a*) 49.25. (*b*) 492.25. (*c*) 492.75. **7.** \$227.53.

Section 95

1. (*a*) $y = 3x$. (*b*) 18. (*c*) y is tripled. **3.** (*a*) $r = v^2/6$. (*b*) 41667 ft. (*c*) 600 ft. per sec. **5.** 95.2 lbs. **11.** 2670 degrees. **13.** (*a*) 50. (*b*) 27.8.

Section 96

7. 3; between 1.4 and 1.5; between -1.4 and -1.5. **9.** between -2.6 and -2.7. **11.** between 1.2 and 1.3. **13.** between 1.2 and 1.3.

Section 100

1. 2.1, 2.01, 2.001, 2.0001, 2.00001. **3.** 1/2, 3/4, 7/8, 15/16, 31/32. **5.** $-1/2$, 1/4, $-1/8$, 1/16, $-1/32$. **7.** 1, 1/2, 1/3, 1/4, 1/5. **9.** 3/2, 5/4, 9/8, 17/16, 33/32.

Section 101

1. 2. **3.** 1. **5.** 0. **7.** 0. **9.** 1.

Section 102

1. 1. **3.** 0. **5.** $3x^2$. **7.** $2x$. **9.** $8x + 3$. **11.** $-1/x^2$. **13.** $-2/x^3$.

Section 105

1. (*a*) $6x$. (*b*) 12. (*c*) $y = 12x - 12$. **3.** (*a*) x. (*b*) 2. (*c*) $y = 2x - 2$. **5.** (*a*) $4x$. (*b*) 8. (*c*) $y = 8x - 3$. **7.** (*a*) $4x - 3$. (*b*) 5. (*c*) $y = 5x - 3$. **9.** (*a*) $-2/x^3$. (*b*) $-1/4$. (*c*) $x + 4y = 3$. **11.** (*a*) 2. (*b*) 2. (*c*) $y = 2x + 3$. **13.** (*a*) 1. (*b*) 1. (*c*) $y = x$. **15.** (*a*) $2cx + k$.

Section 106

1. (*a*) 64, 32, 0, -32 ft. per sec. (*b*) -32 ft. per sec. per sec.
3. (*a*) $v = 100 - 20t$; $a = -20$. (*b*) 0, -100, -200 ft. per sec.
5. $2\pi r$.

Section 107

1. $12x^2 + 12x$. **3.** $2x - 5$. **5.** $20x^3 + 6x^2 - 6x - 4$. **7.** $6x^2 - 6/x^4$. **9.** $2 - 8/x^3$. **11.** (*a*) 3. (*b*) $y = 3x - 2$. **13.** (*a*) 0. (*b*) $y = 2$. **15.** (*a*) 6.

(*b*) $y = 6x + 18$. **17.** (*a*) -3. (*b*) $3x + y + 12 = 0$. **19.** (*a*) 1/4. (*b*) $4y = x + 3$. **21.** 3, -2. **23.** (*a*) $6x^5$. (*b*) $2x$. (*c*) $4x^3$. (*d*) No. **25.** (*a*) $3x^2$. (*b*) $12x$.

Section 108

1. (3, -7) min. **3.** (2, 7) max. **5.** (-1, 1) max.; (4, -124) min. **7.** None.

Section 109

1. 25 ft. by 25 ft.; 625 sq. ft. **3.** 3 seconds; 144 ft. **5.** 4 ft. by 4 ft. by 4 ft. **7.** 6 seconds; 582 ft. **9.** 24 cm. by 32 cm. **11.** 20 in. by 10 in. by 20/3 in.; 4000/3 cu. in. **13.** \$3.00; \$9000.

Section 111

1. $\dfrac{x^2}{2} + 3x + C$. **3.** $5x + C$. **5.** $9x - 2x^3 + C$. **7.** $\dfrac{ax^2}{2} + bx + C$.
9. $96t - 16t^2 + C$.

Section 112

1. $y = 3x + 1$. **3.** $y = x^2 - 4$. **5.** $y = x^2 - 3x + 5$. **7.** 80 ft. per sec. **9.** 80 ft. per sec.

Section 113

1. $34\frac{1}{2}$ square units. **3.** $10\frac{1}{2}$ square units. **5.** 21 square units.

Section 115

1. .7833.

Section 118

1. $\pi/2$. **3.** $\pi/3$. **5.** $\pi/18$. **7.** 180°. **9.** 30°. **11.** $\left(\dfrac{270}{\pi}\right)^\circ$ or 85.94°.
13. $\left(\dfrac{90}{\pi}\right)^\circ$ or 28.65°.

Section 120

Where all six functions are required, they are given in the order sin, cos, tan, csc sec, cot.

1. 3/5, 4/5, 3/4, 5/3, 5/4, 4/3. **3.** 5/13, 12/13, 5/12, 13/5, 13/12, 12/5. **5.** $4/\sqrt{41}$, $5/\sqrt{41}$, 4/5, $\sqrt{41}/4$, $\sqrt{41}/5$, 5/4. **7.** $\sqrt{3}/\sqrt{19}$, $4/\sqrt{19}$, $\sqrt{3}/4$, $\sqrt{19}/\sqrt{3}$, $\sqrt{19}/4$, $4/\sqrt{3}$. **9.** 1/2, $\sqrt{3}/2$, $1/\sqrt{3}$, 2, $2/\sqrt{3}$, $\sqrt{3}$. **11.** 45°. **13.** 60°. **15.** 30°. **17.** 3/5, 4/5, 3/4, 5/3, 5/4, 4/3. **19.** 12/13, 5/13, 12/5, 13/12, 13/5, 5/12. **21.** $\sqrt{7}/4$, 3/4, $\sqrt{7}/3$, $4/\sqrt{7}$, 4/3, $3/\sqrt{7}$. **23.** 4/5, 3/5, 4/3, 5/4, 5/3, 3/4. **25.** $5/\sqrt{34}$, $3/\sqrt{34}$, 5/3, $\sqrt{34}/5$, $\sqrt{34}/3$, 3/5. **27.** 3/5, 4/5, 3/4, 5/3, 5/4, 4/3. **29.** 4/5, 3/5, 4/3, 5/4, 5/3, 3/4. **31.** $6/\sqrt{61}$, $5/\sqrt{61}$, 6/5, $\sqrt{61}/6$, $\sqrt{61}/5$, 5/6. **33.** $2\sqrt{6}/7$, 5/7, $2\sqrt{6}/5$, $7/2\sqrt{6}$, 7/5, $5/2\sqrt{6}$. **35.** (*a*) 1/3. (*b*) $2\sqrt{2}/3$. (*c*) $2\sqrt{2}/3$. **37.** .454. **39.** 1.39.

Section 121

1. 105.0 ft. **3.** 46°. **5.** 67.3 ft. **7.** 235.6 ft. **9.** (*a*) 644.2 ft. (*b*) 145.1 ft. **11.** 4770.3 ft. **13.** 18,480 mi.; 1848 mi. **15.** 238,698 mi. **17.** 66°.

Section 122

1. 125 lbs.; 53° north of east. **3.** $100\sqrt{2}$ lbs. up the plane. **5.** 7.0 mi. per hr.; 79.7 m.p.h. **7.** Any force greater than 845.2 lbs., up the ramp. **9.** 39°.

Section 123

1. 370°, 730°, 1090°, −350°, −710°. **3.** 420°, 780°, 1140°, −300°, −660°. **5.** 570°, 930°, 1290°, −150°, −510°.

Section 124

Where all six functions are required, they are given in the order sin, cos, tan, csc, sec, cot.

1. 4/5, 3/5, 4/3, 5/4, 5/3, 3/4. **3.** −4/5, −3/5, 4/3, −5/4, −5/3, 3/4. **5.** −3/5, 4/5, −3/4, −5/3, 5/4, −4/3. **7.** 0,1,0, does not exist, 1, does not exist. **9.** 12/13, 5/13, 12/5, 13/12, 13/5, 5/12. **11.** 1/2, $\sqrt{3}/2$, $1/\sqrt{3}$, 2, $2/\sqrt{3}$, $\sqrt{3}$. **13.** $-\sqrt{3}/2$, −1/2, $\sqrt{3}$, $-2/\sqrt{3}$, −2, $1/\sqrt{3}$. **15.** 1/2, $-\sqrt{3}/2$, $-1/\sqrt{3}$, 2, $-2/\sqrt{3}$, $-\sqrt{3}$. **17.** Same as ex. 13. **19.** $\sqrt{3}/2$, 1/2, $\sqrt{3}$, $2/\sqrt{3}$, 2, $1/\sqrt{3}$. **21.** $1/\sqrt{2}$, $1/\sqrt{2}$, 1, $\sqrt{2}$, $\sqrt{2}$, 1. **23.** $-1/\sqrt{2}$, $-1/\sqrt{2}$, 1, $-\sqrt{2}$, $-\sqrt{2}$, 1. **25.** Same as ex. 21. **27.** −1, 0, does not exist, −1, does not exist, 0. **29.** 0, 1, 0, does not exist, 1, does not exist. **31.** II, III. **33.** III, IV. **35.** II, IV. **37.** 3/4, $-\sqrt{7}/4$, $-3/\sqrt{7}$, 4/3, $-4/\sqrt{7}$, $-\sqrt{7}/3$. **39.** $-3/\sqrt{34}$, $-5/\sqrt{34}$, 3/5, $-\sqrt{34}/3$, $-\sqrt{34}/5$, 5/3.

Section 126

Where all six functions are required, they are given in the order sin, cos, tan, csc, sec, cot.

1. .1736, −.9848, −.1763, 5.759, −1.015, −5.671. **3.** −.1736, .9848, −.1763, −5.759, 1.015, −5.671. **5.** Same as ex. 1. **7.** .2588, .9659, .2679, 3.864, 1.035, 3.732. **9.** −.6018, −.7986, .7536, −1.662, −1.252, 1.327. **11.** −.2588, −.9659, .2679, −3.864, −1.035, 3.732. **15.** 147°. **17.** 120°. **19.** 147°. **21.** 150°. **23.** 110°.

Section 128

1. $C = 65°$, $a = 11.1$ ft., $b = 15.8$ ft. **3.** $B = 29°$, $b = 12.4$ ft., $c = 19.1$ ft. **5.** 83.3 yds. **7.** (*a*) 7160 ft. (*b*) 7439 ft. (*c*) 6847 ft. **9.** 1165 ft.

Section 129

1. $A = 38°$, $B = 22°$, $C = 120°$. **3.** $c = 21$, $A = 38°$, $B = 22°$. **5.** $b = 50.4$, $A = 55°$, $C = 46°$. **7.** $a = 25.6$, $B = 23°$, $C = 34°$. **9.** 35°, 86°, 59°. **11.** 2371 yds. **13.** 70 lbs.; 38° with 30 lb. force; 22° with 50 lb. force. **15.** 411 yds.

Section 132

3. 1/36. **5.** 1/2; 3/4; 1/4; 1/4. **7.** 3/8; 3/8; 1/8; 1/8; 7/8; 1/2. **9.** 1/4; 1/13; 1/52.

Section 139

1. (*a*) 80. (*b*) 81.5. (*c*) 85. (*d*) 85–89. (*e*) 70–79. (*f*) 84. (*g*) 8.

3. (*a*) 84. (*b*) 87.5 (*c*) none. (*d*) 85–89. (*e*) 90–100. (*f*) 88.3. (*g*) 10.

Section 144

1. (*a*) 20. (*b*) 300. (*c*) $4k$. (*d*) $4(k + 1)$. (*e*) $4(k + 6)$.

3. (*a*) 9. (*b*) 149. (*c*) $2k - 1$. (*d*) $2k + 1$. (*e*) $2k + 11$.

5. (*a*) 15. (*b*) 225. (*c*) $3k$. (*d*) $3(k + 1)$. (*e*) $3(k + 6)$.

Section 162

1. $1/\sin A$; $1/\cos A$; $\sin A/\cos A$; $\cos A/\sin A$.

Section 163

5. 63/65. **7.** −63/16. **9.** 56/65. **11.** 63/65. **13.** 63/16. **15.** 56/65. **17.** 63/65. **19.** −63/16. **21.** −56/65. **23.** 33/65. **25.** −33/56. **27.** 16/65.

Section 164

1. $\frac{1}{2}\sqrt{2-\sqrt{3}}$; $\frac{1}{2}\sqrt{2+\sqrt{3}}$; $2-\sqrt{3}$. **3.** 24/25. **5.** −24/7. **7.** $2/\sqrt{5}$. **9.** −24/25. **11.** 24/7. **13.** $1/\sqrt{5}$. **17.** $\frac{1}{2}\sqrt{2+\sqrt{2}}$; $\frac{1}{2}\sqrt{2-\sqrt{2}}$; $\sqrt{2}+1$.

Section 165

1. 30° + n360°, 150° + n360°; 30°, −330°, 150°, −210°; 30°.
3. 60° + n360°, 300° + n360°; 60°, −300°, 300°, −60°; 60°.
5. 45° + n360°, 225° + n360°; 45°, −315°, 225°, −135°; 45°.
7. 60° + n360°, 120° + n360°; 60°, −300°, 120°, −240°; 60°.
9. 60° + n360°, 240° + n360°; 60°, −300°, 240°, −120°; 60°.
11. 135° + n360°, 225° + n360°; 135°, −225°, 225°, −135°; 135°.
13. 45° + n360°, 315° + n360°; 45°, −315°, 315°, −45°; 45°.
15. 3/4. **17.** $2\sqrt{2}/3$. **19.** 5/4. **21.** 1/2.

Section 166

1. 30° + n360°, 150° + n360°; 30°, −330°, 150°, −210°.
3. 30° + n360°, 150° + n360°, 210° + n360°, 330° + n360°; 30°, −330°, 150°, −210°, 210°, −150°, 330°, −30°.
5. 30° + n360°, 150° + n360°, 270° + n360°; 30°, −330°, 150°, −210°, 270°, −90°.
7. 120° + n360°, 240° + n360°, 0° + n360°; 120°, −240°, 240°, −120°, 0°, 360°, −360°.
9. 90° + n360°, 270° + n360°, 30° + n360°, 150° + n360°; 90°, −270°, 270°, −90°, 30°, −330°, 150°, −210°.
11. same as ex. 9.
13. 0° + n360°, 180° + n360°, 120° + n360°, 240° + n360°; 0°, 360°, −360°, 180°, −180°, 120°, −240°, 240°, −120°.
15. 0° + n360°, 180° + n360°, 30° + n360°, 150° + n360°; 0°, 360°, −360°, 180°, −180°, 30°, −330°, 150°, −210°.
17. 0° + n360°, 180° + n360°; 0°, 360°, −360°, 180°, −180°.
19. 231° + n360°, 309° + n360°; 231°, −129°, 309°, −51°.
21. 15° + n180°, 75° + n180°; 15°, 195°, −165°, −345°, 75°, 255°, −105°, −285°.

Section 167

1. B = 46°, C = 99°, c = 13.8; B = 134°, C = 11°, c = 2.7.
3. No solution.
5. B = 25°, C = 125°, c = 19.7.
7. B = 30°, C = 120°, c = 17.3.
9. 27.6 million mi., 147.2 million mi.

Section 168

(b) **1.** (3, −4), 5.
(b) **3.** (1, −3), 4.

Section 169

1. 1. **3.** −1. **5.** 20. **7.** −21. **9.** True. **11.** True. **13.** False. **15.** True. **17.** False. **19.** False. **21.** True.

Section 170

1. 1, 3, −2. **3.** 1, 6, −2. **5.** 1, 5, −2, −3. **7.** 2/3, −1/3, $\pm\sqrt{-1}$. **9.** −1/2, −1/2, $\frac{1\pm\sqrt{5}}{2}$. **11.** None. **13.** 3, 3, $\pm\sqrt{-1}$.

Section 171

1. 720. **3.** 72. **5.** 72. **7.** 120. **9.** 20. **11.** 60. **13.** (*a*) 604,800. (*b*) 10,000,000. **15.** 6840. **17.** 216. **19.** (*a*) 48. (*b*) 72.

Section 172

1. 10. **3.** 190. **5.** 120. **7.** 35. **9.** 350. **11.** (*a*) 112. (*b*) 140. **13.** (*a*) 45. (*b*) 9. **15.** 403,200. **17.** (*a*) 3003. (*b*) 1001. (*c*) 2002.

Section 173

1. $x^7 + 7\,x^6h + 21\,x^5h^2 + 35\,x^4h^3 + 35\,x^3h^4 + 21\,x^2h^5 + 7\,xh^6 + h^7$.

3. $p^9 + 9\ p^8q + 36\ p^7q^2 + 84\ p^6q^3 + 126\ p^5q^4 + 126\ p^4q^5 + 84\ p^3q^6 + 36\ p^2q^7 + 9\,pq^8 + q^9$.

5. $32\,a^{15} + 240\,a^{12}b^2 + 720\,a^9b^4 + 1080\,a^6b^6 + 810\,a^3b^8 + 243\,b^{10}$.

7. $a^3 - 3\,a^2b + 3\,ab^2 - b^3$. **9.** $a^5 - 5\,a^4b + 10\,a^3b^2 - 10\,a^2b^3 + 5\,ab^4 - b^5$.

11. 9/4. **13.** 625/256. **15.** $680\,x^{14}h^3$. **17.** $3003\,x^{10}h^5$. **19.** $\frac{15}{2}\,a^{14}b^6$.

21. $70\,a^4b^4$. **23.** $84\,b^3/a^6$. **25.** (*a*) 1.094. (*b*) .942. (*c*) 1.242. (*d*) .868.

Section 174

1. 1/36. **3.** (*a*) 1/28561. (*b*) 1/270725. **5.** 3/13. **7.** (*a*) 3/11. (*b*) 2/33. **9.** 625/3888. **11.** (*a*) 15/128. (*b*) 121/128. **13.** (*a*) 125/3888. (*b*) 69/1944. (*c*) 3875/3888. **15.** (*a*) 1/7776. (*b*) 1/60466176. (*c*) No.

INDEX